Fact Checking with Better Retrieval: Dynamic Adaptive Contrastive Learning for Evidence Retrieval

Anonymous ACL submission

Abstract

In the field of multimodal fact checking, the accuracy of retrieving evidence from different modalities has a significant impact on the downstream claim verification process.The existing methods usually rely on semantic feature extraction. The retriever tends to focus on the semantic-rich parts of the features and ignores the features at the event skeleton level, resulting in the positive evidence ranking lower in the retrieval results. This paper proposes a method called **DACLR** to address the problem. DACLR first uses a Multimodal Large Language Model (MLLM) to enhance the event information of multimodal claims and evidence. Then, it conducts evidence retrieval through a two-stage retrieval method of recall-rerank. DACLR enhances the model's event perception ability of the retrieval stage by optimizing the contrastive loss and mining hard negative samples. Specifically, DACLR designs three loss functions and three sets of negative sample candidate at two levels (semantic and event). The model dynamically adjusts the proportion of loss functions and negative samples based on the self-supervised signals during training, enabling the model to learn the correlations between samples at the event level without losing its semantic capability. Comprehensive comparison and ablation experiments demonstrate the effectiveness of DACLR.

1 Introduction

Multimodal fact checking aims to search for and integrate multimodal evidence (e.g., text, tables, images) from the specified multimodal evidence database to verify the veracity of the given text(Yang et al., 2018). Current research(Zhou et al., 2019) typically follows a two-stage paradigm of evidence retrieval and claim verification. In the evidence retrieval stage, fine-grained evidence related to the claim is retrieved from a large-scale evidence database(Guo et al., 2022b). In the claim verification stage, the authenticity of the claim is determined based on the retrieved evidence(Wu et al., 2021). Existing studies(Zheng et al., 2024) have demonstrated the significance of the evidence selected in the retrieval stage for the entire process. The more accurate the information related to the claim in the retrieved evidence is, the more accurate the judgment made by the verification system will be.

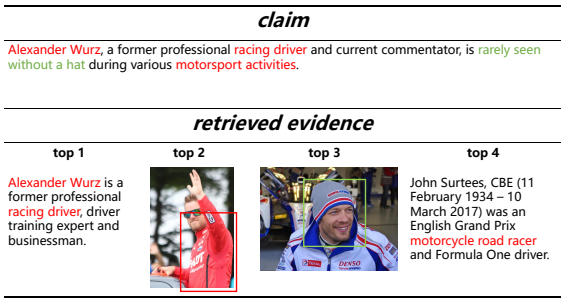

Figure 1: An example of semantic-based multimodal evidence retrieval: Although the first two pieces of evidence are highly relevant to the claim at the semantic level, the evidence that provides the most relevant information is placed at the lower position.

Although the current retrieval solutions for the field of fact checking have made significant progress(Zhou et al., 2019; Zheng et al., 2024; Wu et al., 2021), most of the methods focus on text modal retrieval. On the other hand, multimodal retrieval solutions(Zeng et al., 2025; Zhou et al., 2024) for general domains usually construct the retrieval scheme based on semantic similarity. In this scheme, all candidate multimodal evidence and claims are mapped to the same latent semantic space by an embedding model, and then the top k candidate evidence is selected based on the similarity between the features. One common issue that arises in the practical application of these methods is that the evidence providing the most relevant information for claim verification does not rank high

enough, resulting in insufficient information in the subsequent verification stage. Figure 1 shows the 4 results returned by the semantic retrieval model for the claim to be verified. It can be seen that the first two pieces of evidence are highly similar in semantics to the claim, but both lack the key fact of "wearing a hat" at the event level, while the evidence truly containing this event ranks much lower. Thus, although these models may perform well in general multimodal retrieval, they are difficult to directly handle the task of fact checking.The fundamental cause of the aforementioned problem lies in the fact that the model performs feature extraction of multimodal information during training and conducts semantic learning based on this. Semantic feature extraction often relies on semantic feature information such as participants and attributes, which tends to overlook the feature at the event skeleton level, thereby resulting in retrieval results that are similar but not relevant(Zhao et al., 2023).

To address the aforementioned challenges, this paper proposes a multimodal fact checking retrieval method based on dynamic adaptive contrastive learning, named DACLR (**D**ynamic **A**daptive **C**ontrastive **L**earning for **R**etrieval). DACLR is divided into three main parts: event summary generation, two-stage evidence retrieval, and model training optimization. In the event summary generation part, to map claims and evidence to a unified event level, DACLR uses a predefined prompt to guide the Multimodal Large Language Model(MLLM) to generate event summaries containing event skeletons, participants, and attributes for claims and evidence. In the two-stage evidence retrieval, the bi-encoder(Guo et al., 2022a) is used in the recall process to improve the computational efficiency, while the cross-encoder(Zhang et al., 2022a) is used in the re-ranking stage to obtain better results.In the model training optimization, to enhance the model's ability to capture the association between claim and evidence at event skeleton level, DACLR designs a dynamic adaptive contrastive learning method. This method uses the in-batch accuracy during the training as the supervisory signal, dynamically adjusting the loss functions at the event level and semantic level, as well as the negative samples with high difficulty. Through this adaptive mechanism based on the model's own training state, DACLR can effectively learn the features at the event level while retaining the knowledge at the semantic level.

In summary, the contributions of this paper can be summarized as:(i)A multimodal evidence retrieval method named DACLR for fact checking is proposed. Through MLLM, the multimodal data is enhanced with event information, enabling the retriever to learn the feature connections between samples at the event skeleton level, thereby improving the ranking of positive sample evidence in the retrieval results.(ii)In order to enable the retriever to better utilize the enhanced event information for learning, a dynamic adaptive contrastive learning method is designed. Through the self-supervised signals of the model during training, the ratio of the loss function and the distribution of negative samples are dynamically adjusted.(iii)Experiments are conducted to prove the effectiveness of DACLR. Detailed experimental analysis shows that DACLR not only performs better than existing multimodal retrieval methods in the multimodal datasets, but also outperforms existing fact checking retrieval methods in the single-modal datasets.

2 Related Work

2.1 Retrieval for fact checking

The mainstream paradigm of fact checking can be divided into two parts: retrieval and verification. The evidence retrieval results obtained in the retrieval stage have a significant impact on the subsequent verification. The early retrieval relied on TFIDF and BM25 (Robertson and Zaragoza, 2009), which encode claims and evidence into sparse bag-of-words vectors to achieve efficient word matching. However, these methods are limited by the vocabulary gap problem and are difficult to capture the semantic relationship between claims and evidence. After the emergence of transformer(Vaswani et al., 2017) and its variants, retrieval systems based on pre-trained models have become the mainstream paradigm(Humeau et al., 2020). This method encodes queries and documents into dense semantic vectors (Ren et al., 2021; Zhang et al., 2022b), and uses an optimized similarity search algorithm(Johnson et al., 2021) to perform retrieval. These methods require training a retrieval model using a contrastive learning paradigm or a supervised fine-tuning paradigm, thereby obtaining the evidence with the highest similarity to the claim.

However, in the field of checking, the evidence that is most similar to a claim might not be the most relevant one, and the truly relevant evidence might be ranked lower by the retrieval model due

to insufficient semantic similarity. To address this issue, some studies(Zheng et al., 2024) have attempted to achieve joint optimization between the retriever and the verification model by introducing KL divergence as a supervisory signal. Other studies(Malviya and Katsigiannis, 2024) have utilized the interaction relationships between evidences through multi-stage reordering. However, these methods still focus on semantic similarity in the retrieval stage, failing to fundamentally solve the retrieval bias problems of the model at the semantic and event levels. Moreover, most of the existing research is focused on pure text modal fact checking, and there is an urgent need for a unified retrieval framework for multimodal fact checking.

2.2 Multimodal retrieval

Compared with text-modal retrieval, multimodal retrieval requires the system to capture the features and connections between information from different modalities (Yuan et al., 2021). A mainstream approach is to use representation models. Early studies encoded the information of the visual and the text information separately and concatenated the vectors for information fusion (Radford et al., 2021). Recent research (Li et al., 2024b) focuses on projection-based methods. This type of method utilizes models like CLIP (Radford et al., 2021) to convert visual inputs into feature sequences and introduces a projection layer to align these sequences with the text embeddings. For example, MARVEL(Zhou et al., 2024) uses this framework to capture multimodal information in the output space of the language model; MCL (Li et al., 2024b) adds a retrieval token, and CIEA(Zeng et al., 2025) captures supplementary information between text and images for better multimodal retrieval. Although these representation model-based methods can achieve effective retrieval in general domains, they are difficult to be adjusted according to the characteristics of specific downstream tasks (such as fact checking).

Another approach is to uniformly convert the information of other modalities into text modality, thereby transforming the multimodal retrieval task into text retrieval task (Baldrati et al., 2023). Although this approach may result in information loss during the modal conversion process (Che et al., 2024; Li et al., 2024c), it is relatively easy to adapt and enhance according to the characteristics of the downstream task. The DACLR follows this method route and has built an end-to-end multimodal retrieval method for fact checking.

3 Method

This section provides a detailed introduction to the proposed method Dynamic Adaptive Contrastive Learning for Retrieval(DACLR). The overall architecture of the method is shown in Figure 2, which consists of three main components: Event Summary Generation, Two-stage Retrieval, and Training Optimization. Given a claim c to be verified and an evidence set $E = \{e_1, e_2, \ldots\ldots, e_n\}$, the goal of DACLR is to retrieve the top q most relevant evidence from the set E that are related to the claim.

3.1 Event Summary Generation

To transform multimodal evidence into a unified representation, MLLMs are employed to extract event summaries for each claim and multimodal evidence. In order to enhance the information alignment between evidence and claims, thereby improving the modeling of factual consistency and the sorting of retrieval results. Drawing on the discourse semantic theory (Penn Discourse Treebank, PDTB)(Prasad et al., 2008), the claim/evidence is decomposed into structure, participants and attributes.Specifically, for the input x, the model outputs an event skeleton tuple:

$$S(x) = \{summary(x), participants(x), attributes(x), structure(x)\} \quad (1)$$

where summary is a natural language summary of an event, including participants, attributes, and event skeleton, participants and attributes respectively list the participant entities and their attributes in the event, structure is the remaining event skeleton after masking all the participants and attributes in the event summary with masks, such as '[Mask] is involved in [Mask]'. The prompt used in this section is presented in the Appendix E.

3.2 Two-stage Retrieval

During the retrieval stage, the classic recall-rerank framework is followed. Among them, the objective of the recall stage is to narrow down the range of n candidate evidences to p (where $q < p < n$). At this stage, the claim and evidence will be independently encoded by the bidirectional encoder of DACLR, and the correlation between vectors will be calculated through cosine similarity. The top p

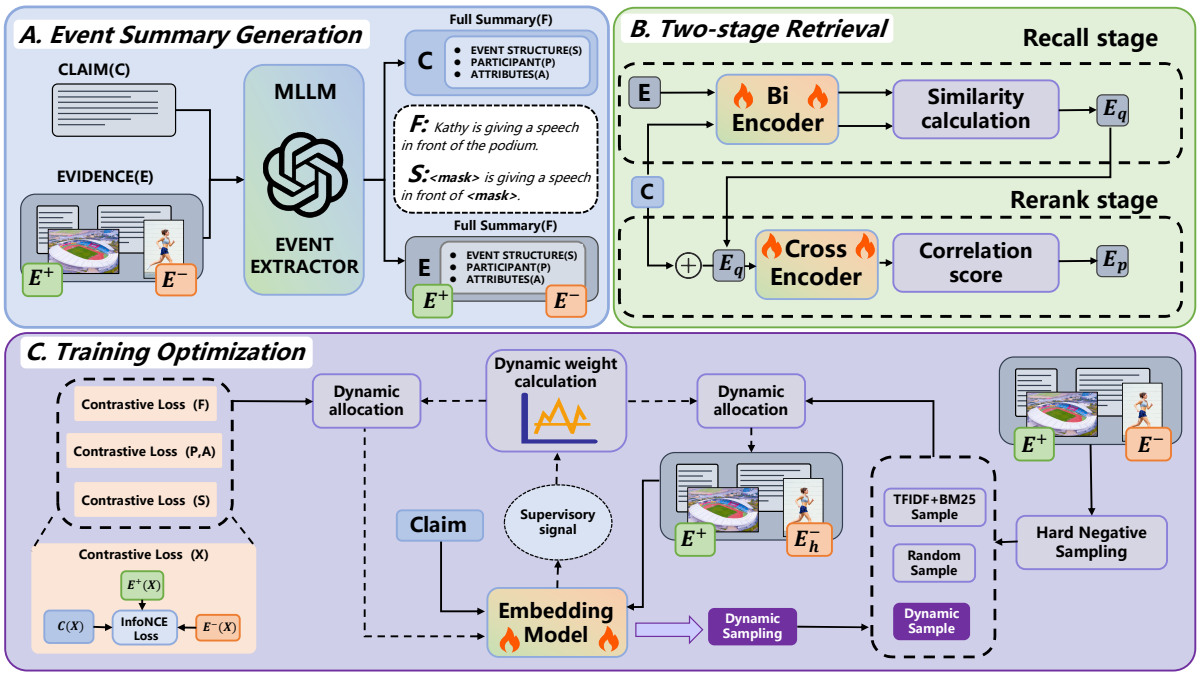

Figure 2: An overview of DACLR.

pieces of evidence with the highest scores are selected. Given the claim and set E, the bidirectional Encoder will encode them to obtain their respective representations:

$$f_c, f_e = Encoder(c, e) \tag{2}$$

Next, for each claim, calculate the pairwise cosine similarity Sn between it and all the evidence in the evidence set:

$$S_n = \{s_1, s_2, \ldots\ldots, s_n\} = \{cos(f_c, f_{e_1}), cos(f_c, f_{e_2}), \ldots\ldots cos(f_c, f_{e_n})\} \tag{3}$$

Sort S_n, and the top p pieces of evidence will be input into the rerank stage as the final result s_p of the recall stage.

The goal of the rerank stage is to select the most relevant q pieces of evidence from the p pieces of s_p. In this stage, DACLR refers to the implementation of the cross-encoder and concatenates the claim and evidence together for feature encoding, thereby obtaining the joint representation of the claim and the evidence. A fully connected layer is used to obtain the score for each piece of evidence in the s_p:

$$\widetilde{S}_p = \{s_1, \ldots\ldots, s_p\} = \{FC(Encoder(c \oplus e_1)), \ldots\ldots, FC(Encoder(c \oplus e_p))\} \tag{4}$$

Sort S_p and ultimately select q pieces of evidence as the final results.

3.3 Training Optimization

In order to enable encoder model to better learn the relationship between claims and positive sample evidence at the event level, DACLR uses a dynamic adaptive contrastive learning method. This method optimizes the two major elements of contrastive learning: contrastive loss and hard negative sample mining, and uses the in-batch accuracy of the model as a self supervised signal to dynamically regulate the training process.

3.3.1 Dynamic loss allocation

DACLR designs two layers of contrastive loss, namely the event summary loss and the semantic-event joint loss. Among them, the event summary loss is calculated using the event summaries, which includes information from both the semantic-level and event-level. Therefore, it participates in the

entire training process and holds a dominant position in the training. The semantic-event joint loss is dynamically matched with the proportion of semantic(participants and attributes) and event(skeletons) in the training based on the self supervised signal of the model.

Event Summary Loss: This loss is calculated using the event summary of claim and evidence. For each claim, calculate the comparative loss:

$$L_{full} = -\log \frac{e^{sim(f_{c_i}^{full}, f_{e^+}^{full})/\tau)}}{e^{sim(f_c^{full}, f_{e^+}^{full})/\tau)} + \sum_{D^-} e^{sim(f_c^{full}, f_{e^-}^{full})/\tau)}} \quad (5)$$

where f^{full} is the representation after encoding and pooling, τ is the temperature hyperparameter, and $D-$ is the currently negative sample set.

semantic-event joint loss: Furthermore, DACLR designs two sub-losses, which are combined into semantic-event joint loss by dynamic parameter adjustment: Semantic loss L_{sent}: Compare the claim's participants and attributes with the corresponding version of the evidence. Event comparison loss L_{struct}: Use the structure field for matching, forcing the model to focus on the structural pattern of the event. The final semantic-event joint loss is:

$$L_{unit} = \beta \cdot L_{sent} + (1 - \beta) \cdot L_{struct} \quad (6)$$

where β is a dynamic weight used to balance the contrastive loss of two layers. The calculation method of this weight is detailed in section 3.3.3.

3.3.2 Dynamic Negative Sample Mining

Traditional contrastive learning usually employs random negative samples or static hard negative samples, but it is difficult to train a retriever with strong discriminative ability. DACLR proposes a hybrid negative sample mining mechanism. The negative sample pool is composed of three parts: **(i).Random Negative Sample** D_{rand}: Selected randomly from the evidence database.**(ii).Shallow Hard Negative Sample** D_{TB}: This sample is based on the integrated retrieval results of TF-IDF and BM25, and gathers evidence that is similar to the claims' words but has a relatively weak semantic correlation. **(iii).Self-Retrieval Hard Negative Sample** D_{model}: Utilizing the most similar but incorrect evidence dynamically mined by the current retrieval model during training, these samples are constantly updated as the model's capability improves. DACLR refers to existing research(Li et al., 2024a) and updates based on the model's score on the negative samples.

During the training process, for each claim, the total number of negative samples is fixed at K. Among them, the proportion of fixed negative samples (D_{rand} and D_{TB}) is $1 - p_{dyn}$, and the proportion of dynamic negative samples is pdyn. The calculation method of the proportion p_{dyn} is elaborated in Section 3.3.3.

3.3.3 Adaptive Scheduling based on Discrimination

DACLR designs an adaptive scheduling mechanism based on the current discrimination ability of the model to calculate the matching parameters β and p_{dyn}. The average positive and negative sample similarity gap of each batch is defined as:

$$\Delta_t = \frac{1}{B} \sum_{i \in B} \Big(\mathrm{sim}(f_{c_i}^{\mathrm{full}}, f_{e_i^+}^{\mathrm{full}}) - \frac{1}{D} \sum_{j \in D} \mathrm{sim}(f_{c_i}^{\mathrm{full}}, f_{e_j^-}^{\mathrm{full}}) \Big) \quad (7)$$

And use the exponential moving average method to smooth and update Δ_t:

$$\bar{\Delta}_t = 0.9\bar{\Delta}_{t-1} + 0.1\Delta_t \quad (8)$$

In order to smoothly map the Δ_t to a certain value within the (0, 1) interval, DACLR uses the sigmoid function for mapping, and periodically calculates the accuracy ACC_{val} on the validation set, using it as the median threshold Δ_{mid} for the sigmoid function. Finally, the dynamic negative sample ratio and the loss weight are calculated as follows:

$$p_{dyn} = \sigma(\frac{\bar{\Delta}_t - \Delta_{mid}}{\tau}), \beta = 1 - p_{dyn} \quad (9)$$

Where σ is the sigmoid function, and τ controls the steepness of the transition. When the model's discrimination ability is weak, it mainly uses simple negative samples and focuses on semantic feature learning; when the model gradually becomes stronger, it automatically introduces more difficult negative samples and shifts to event skeleton learning. The pseudo-code of the training process is presented in Appendix A.

Modality	Method	FEVER						FEVEROUS					
		REC@20	REC@100	MRR@10	MRR@20	NDCG@10	NDCG@20	REC@20	REC@100	MRR@10	MRR@20	NDCG@10	NDCG@20
Text-only	TF-IDF	47.87	70.26	31.74	33.57	28.99	29.62	67.11	71.56	29.71	29.89	30.14	29.93
	BM25	46.34	68.77	28.02	28.32	30.23	31.26	69.40	75.53	33.71	34.03	31.05	31.74
	BERT	87.47	90.34	57.23	58.36	58.12	59.47	81.54	**89.33**	54.17	55.93	52.49	53.61
	XLNet	85.34	88.46	60.23	61.92	62.87	64.14	76.32	82.95	58.12	59.73	55.42	56.21
	RoBERTa	85.57	90.65	55.24	53.67	54.71	55.24	80.35	88.13	55.38	57.42	59.29	60.01
	GERE	77.16	88.75	59.83	61.22	61.72	62.33	84.43	87.62	66.03	66.84	67.09	67.93
	KGAT	87.47	89.63	61.73	63.77	59.88	60.74	86.04	87.35	64.79	66.21	63.05	64.12
	RAV	87.77	90.57	60.24	62.29	62.42	65.71	85.17	88.74	61.37	63.30	64.15	64.24
	M-ReRank	86.13	**90.78**	63.27	63.82	65.34	67.92	**87.02**	88.13	67.21	67.88	65.19	66.23
	DACLR	**88.48**	89.58	**69.04**	**69.32**	**70.28**	**71.29**	86.93	88.87	**70.11**	**70.24**	**68.91**	**69.03**
		MMCV						Mocheg					
		rec@20	rec@100	mrr@10	mrr@20	ndcg@10	ndcg@20	rec@20	rec@100	mrr@10	mrr@20	ndcg@10	ndcg@20
Multimodal	CLIP-DPR	55.37	67.36	40.74	39.50	35.33	36.23	44.88	59.63	31.41	32.85	30.74	32.49
	VinVL-DPR	50.12	66.78	33.75	34.74	34.32	35.66	49.71	60.11	33.41	34.71	31.71	31.84
	UniVL-DP	55.24	68.36	39.63	41.24	41.25	40.27	53.74	60.75	36.21	36.78	37.84	39.23
	T5-ANCE	61.35	69.96	44.25	46.34	46.38	47.25	54.32	63.79	34.77	34.93	36.87	37.01
	MARVEL	63.46	79.77	50.76	50.97	44.35	48.06	61.74	65.29	49.23	50.12	48.13	48.65
	CIEA	67.70	80.47	51.75	52.67	46.73	45.69	64.23	70.14	51.12	51.84	53.73	54.14
	DACLR	**71.16**	**80.83**	**66.26**	**67.13**	**62.43**	**63.91**	**67.43**	**76.87**	**60.16**	**60.67**	**59.41**	**60.12**

Table 1: This table shows the main experimental results.

4 Experiments

4.1 Setup

4.1.1 Dataset and Metrics

To comprehensively evaluate the retrieval capabilities of DACLR in the field of fact checking, following datasets are selected:

Text-only dataset: FEVER(Thorne et al., 2018),FEVEROUS(Aly et al., 2021)

Multimodal dataset: MMCV(Wang et al., 2025),Mocheg(Yao et al., 2023).

The detailed statistics of all datasets is provided in the Appendix B.

In order to compare the performance of the DACLR with that of the baseline methods, Recall@k is used as one of the metrics in the experiment. Additionally, to analyze the ranking of positive samples among all retrieved samples, MRR@k and NDCG@k is also used in the experiment. For the values of k in all the above-mentioned metrics, they are as follows: [10, 20, 100]

4.1.2 Baselines and Details

Two different types of baselines are used to be compared with DACLR:

Evidence retrieval for text modalities:This type of method focuses on selecting the most relevant evidence from a text database to support a claim. For instance, the **BM25/TFIDF** methods adopted by the official FEVER dataset. Recent studies usually conduct dense retrieval based on retrieval models such as **BERT**(Devlin et al., 2019), **XLNet**(Zhong et al., 2020), and **RoBERTa**(Zhong et al., 2020). Among them, **GERE**(Zhou et al., 2019), **KGAT**(Liu et al., 2020), **RAV**(Zheng et al., 2024), and **M-ReRank**(Malviya and Katsigiannis, 2024) are used as representatives of the latest methods.

Multimodal Evidence Retrieval:This type of method aims to find the evidence from a multimodal evidence database. The latest research selected here serves as the baselines : Pre-trained multimodal model VinVL-DPR (Zhang et al., 2021) and CLIP-DPR (Radford et al., 2021), Hard negative samples mining based method UniVL-DR(Liu et al., 2023), Projection-based method MARVEL (Zhou et al., 2024) and CIEA(Zeng et al., 2025).

Implementation Details The RoBERTa model is used as the base model for training.Due to page limitations, please refer to the Appendix C for the implementation details of all algorithms used in the experiment and the corresponding hyperparameters.

4.2 Main Results

Table 1 presents the comparative main results of DACLR and various baselines in the evidence retrieval task. Overall, DACLR achieves significant improvements in retrieval performance compared to existing methods. Although it does not show a clear advantage in the recall, it achieves a substan-

Method	REC@20	MRR@20	NDCG@20
RoBERTa	85.57	53.67	55.24
DACLR w/o Event Loss	87.47	64.12	65.47
DACLR w/o Hard Samples	86.96	63.31	66.12
DACLR	**88.48**	**69.32**	**71.29**

Table 2: This table shows the ablation experimental results on FEVER Dataset.

tial lead in MRR and NDCG. This result is consistent with the design goal of DACLR, which is to improve the ranking quality of the most relevant evidence in the candidate set rather than simply maximizing the number of positive examples.

In the pure **text retrieval** task, DACLR outperforms the recent representative methods RAV and M-ReRank on both the FEVER and FEVEROUS datasets. Specifically, DACLR achieves 69.04 MRR@10 and 70.28 NDCG@10 on the FEVER dataset, and 70.11 and 68.91 respectively on the FEVEROUS dataset. The above experimental results indicate that DACLR can effectively model and retrieve evidence at the event level, thereby elevating positive evidence to a more prominent position in the retrieval list.

In the **multimodal retrieval** task, DACLR also demonstrates good generalization ability. On the MMCV and Mocheg multimodal datasets, this method achieves 66.26 and 60.16 MRR@10, and 62.43 and 59.41 NDCG@10 respectively. Although the introduction of heterogeneous representation alignment difficulties for image and text in multimodal data poses challenges, DACLR can maintain relatively stable performance. The reason for the lower results in the multimodal task compared to the text task may be due to the semantic gap between the visual and text modalities, which increases the difficulty of cross-modal retrieval. Nevertheless, DACLR can effectively coordinate heterogeneous information through the unified representation of event structure, demonstrating good cross-modal retrieval potential.

4.3 Ablation Study

The core of the DACLR design lies in the dynamic allocation of the event loss function and the mining of difficult negative samples. By separately removing the dynamic allocation loss function (i.e., only using the InfoNCE function for training) and the mining of difficult negative samples (i.e., only using random negative samples), the effects and contributions of both are verified on DACLR. The experimental results are shown in Table 2. Overall,

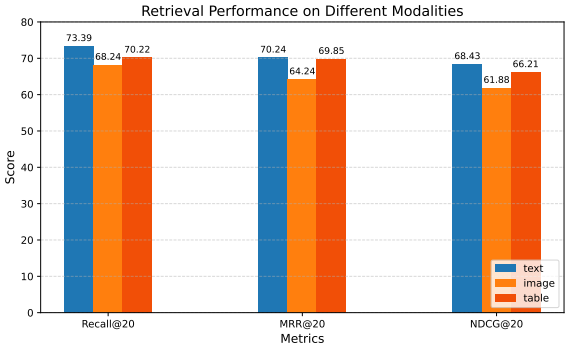

Figure 3: A figure shows the retrieval performance on different Modalities.

using the DACLR method is significantly superior to the RoBERTa model trained with standard contrastive learning in all indicators. Specifically, DACLR outperforms either without the event loss or any of the hard samples, indicating the effectiveness of these two for model training optimization.

Furthermore, in order to comprehensively analyze the retrieval capabilities of DACLR for various modalities of evidence, experiments for each modality of evidence are conducted on the MMCV dataset, including text, images, and tables. The experimental results are shown in Figure 3. From the results, it can be seen that DACLR has the best retrieval performance for text evidence, while its retrieval performance for image evidence is relatively weak. This is consistent with the results of the comparative experiments. Thanks to the powerful semantic understanding ability of the large language model, DACLR's retrieval ability for table evidence has approached that of text evidence. Both have achieved a Recall@20 of 70+%, a MRR@20 of 69+%, and an NDCG@20 of 66+%.

4.4 Performance on claim verification

The original intention of the DACLR is to select the most relevant evidence from the multi-modal evidence candidate set to assist the downstream model in claim verification. Existing studies have proved the importance of the quality of the retrieved evidence for claim verification. Based on this, further experiments are conducted on MMCV dataset to prove that the evidence extracted by DACLR can be better utilized by the downstream model. Using the classic RAG architecture, the Top-10 evidence retrieved by DACLR is used for verification. GPT-4o(OpenAI et al., 2024) and Gemini 1.5 Flash(Team et al., 2024) are selected as base model, and LLaVA-V1.5-7B(Liu et al., 2024), which is an

Method	Precision	Recall	F1
GPT-4o	70.39	65.74	67.99
GPT-4o + MARVEL	70.82	67.25	68.99
GPT-4o + CIEA	71.36	68.93	70.12
GPT-4o + DACLR	**74.88**	**69.43**	**72.05**
GEMINI	71.53	66.12	68.72
GEMINI + MARVEL	72.34	67.82	70.01
GEMINI + CIEA	72.17	68.35	70.21
GEMINI + DACLR	**74.92**	**70.01**	**72.38**
LLaVa	64.93	63.22	64.06
LLaVa + MARVEL	64.97	63.27	64.11
LLaVa + CIEA	65.05	64.12	64.58
LLaVa + DACLR	**65.07**	**64.18**	**64.62**

Table 3: Performance on claim verification.

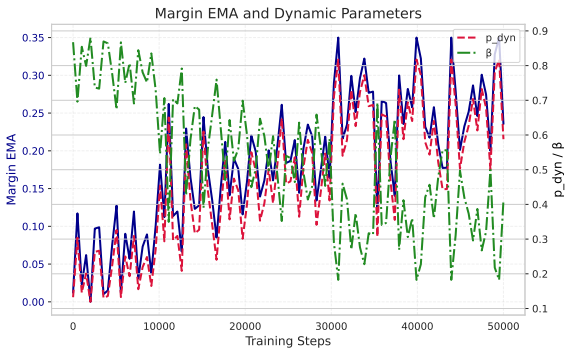

Figure 4: The curve of dynamic parameters during training.

open-source multimodal model representative, is also selected. The experimental results are shown in Table 3. The results indicate that the MLLM using the evidence retrieved by DACLR achieved the best performance indicators. This shows the contribution of DACLR to the entire fact checking process, significantly improving the performance of end-to-end fact checking through better evidence retrieval.

4.5 Dynamic Parameters Analysis

During the DACLR training process, the dynamic parameters p_{dyn} and β represent the ratio of hard negative samples and event loss, and their changing trends during the training process are shown in Figure 4. As the training progresses, p_{dyn} shows a gradually increasing trend, while β is exactly the opposite. Although their values may drop or rise to a certain extent with each negative sample's updating, the overall trend remains unchanged. This also indicates that the model gradually acquires the ability to distinguish positive and negative samples from the perspective of event during the training

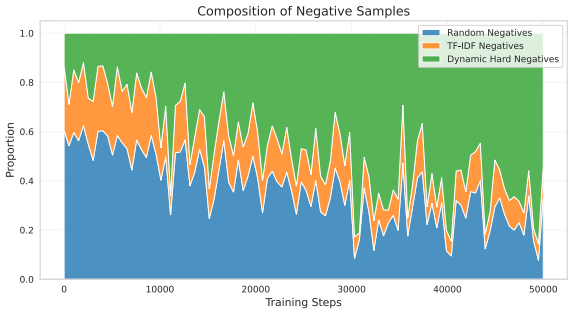

Figure 5: The variation in the proportion of negative samples during training.

process. The proportion distribution of negative samples during the training process is shown in Figure 5. As the model's ability increases, the proportion of difficult negative samples in the negative samples also increases, enabling the model to conduct more detailed learning of positive and negative sample distinctions.

4.6 Conclusion

This paper proposes a multimodal evidence retrieval method named DACLR, which consists of three parts: event summary generation, two-stage evidence retrieval, and model training optimization. Firstly, using MLLM, unified event summaries are extracted for claims and evidence. Then, through the two-stage retrieval, evidence are retrieved according to the claim. In the model training optimization, to enable the model to learn the connection between claims and evidence at the event level, DACLR optimizes the training from two aspects: loss function and negative sample mining. There contrastive losses and hard negative sample sets are designed, and the ratio of the two is dynamically updated based on the intra-batch sample accuracy of the model.

To verify the effectiveness of DACLR, extensive comparative and ablation experiments are conducted. The experimental result demonstrate that DACLR not only outperforms existing general multimodal retrieval methods in multimodal datasets, but also outperforms existing text evidence retrieval methods in pure text evidence retrieval. Moreover, the ablation experiment prove the indispensability of the dynamic loss function and hard negative sample ratio for DACLR. Furthermore, the downstream claim verification experiment also demonstrate that the evidence retrieved by DACLR is more relevant to the claim than that retrieved by existing methods.

Limitations

Although the expected results is achieved, DACLR still has several limitations that need to be addressed in future work. The first limitation lies in its dependence on MLLM. As the first and most crucial step of DACLR, the quality of event summaries generated by MLLM has a significant impact on the final retrieval effect. Although the current understanding and generation capabilities of MLLM are quite strong, its unstable output still needs to be resolved. When the output of MLLM contains poor-quality event summaries, it will have a certain impact on DACLR. Another limitation arising from this is the scalability of the modality. DACLR is mainly focused on text and image modality for fact verification. If it is to be expanded to audio or video modalities in the future, it will be limited by the performance of MLLM in that modality.

References

Rami Aly, Zhijiang Guo, Michael Sejr Schlichtkrull, James Thorne, Andreas Vlachos, Christos Christodoulopoulos, Oana Cocarascu, and Arpit Mittal. 2021. FEVEROUS: Fact extraction and VERification over unstructured and structured information. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*.

Alberto Baldrati, Lorenzo Agnolucci, Marco Bertini, and Alberto Del Bimbo. 2023. Zero-shot composed image retrieval with textual inversion. *Preprint*, arXiv:2303.15247.

Chang Che, Qunwei Lin, Xinyu Zhao, Jiaxin Huang, and Liqiang Yu. 2024. Enhancing multimodal understanding with clip-based image-to-text transformation. *Preprint*, arXiv:2401.06167.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Cynthia Dwork, Ravi Kumar, Moni Naor, and D. Sivakumar. 2001. Rank aggregation methods for the web. In *Proceedings of the 10th International Conference on World Wide Web*, WWW '01, page 613–622, New York, NY, USA. Association for Computing Machinery.

Jiafeng Guo, Yinqiong Cai, Yixing Fan, Fei Sun, Ruqing Zhang, and Xueqi Cheng. 2022a. Semantic models for the first-stage retrieval: A comprehensive review. *ACM Trans. Inf. Syst.*, 40(4).

Zhijiang Guo, Michael Schlichtkrull, and Andreas Vlachos. 2022b. A survey on automated fact-checking. *Transactions of the Association for Computational Linguistics*, 10:178–206.

Samuel Humeau, Kurt Shuster, Marie-Anne Lachaux, and Jason Weston. 2020. Poly-encoders: Transformer architectures and pre-training strategies for fast and accurate multi-sentence scoring. *Preprint*, arXiv:1905.01969.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2021. Billion-scale similarity search with gpus. *IEEE Transactions on Big Data*, 7(3):535–547.

Shiyu Li, Yang Tang, Shizhe Chen, and Xi Chen. 2024a. Conan-embedding: General text embedding with more and better negative samples. *Preprint*, arXiv:2408.15710.

Wei Li, Hehe Fan, Yongkang Wong, Yi Yang, and Mohan Kankanhalli. 2024b. Improving context understanding in multimodal large language models via multimodal composition learning. In *Forty-first International Conference on Machine Learning*.

Zhang Li, Biao Yang, Qiang Liu, Zhiyin Ma, Shuo Zhang, Jingxu Yang, Yabo Sun, Yuliang Liu, and Xiang Bai. 2024c. Monkey: Image resolution and text label are important things for large multi-modal models. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 26753–26763.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024. Improved baselines with visual instruction tuning. *Preprint*, arXiv:2310.03744.

Zhenghao Liu, Chenyan Xiong, Yuanhuiyi Lv, Zhiyuan Liu, and Ge Yu. 2023. Universal vision-language dense retrieval: Learning a unified representation space for multi-modal retrieval. *Preprint*, arXiv:2209.00179.

Zhenghao Liu, Chenyan Xiong, Maosong Sun, and Zhiyuan Liu. 2020. Fine-grained fact verification with kernel graph attention network. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7342–7351, Online. Association for Computational Linguistics.

Shrikant Malviya and Stamos Katsigiannis. 2024. Evidence retrieval for fact verification using multi-stage reranking. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 7295–7308, Miami, Florida, USA. Association for Computational Linguistics.

OpenAI, :, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Mądry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex

Chow, Alex Kirillov, and 401 others. 2024. Gpt-4o system card. *Preprint*, arXiv:2410.21276.

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind Joshi, and Bonnie Webber. 2008. The Penn Discourse TreeBank 2.0. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA).

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. *Preprint*, arXiv:2103.00020.

Ruiyang Ren, Shangwen Lv, Yingqi Qu, Jing Liu, Wayne Xin Zhao, QiaoQiao She, Hua Wu, Haifeng Wang, and Ji-Rong Wen. 2021. PAIR: Leveraging passage-centric similarity relation for improving dense passage retrieval. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2173–2183, Online. Association for Computational Linguistics.

Stephen Robertson and Hugo Zaragoza. 2009. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends in Information Retrieval*, 4(1-2):1–174.

Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, Soroosh Mariooryad, Yifan Ding, Xinyang Geng, Fred Alcober, Roy Frostig, Mark Omernick, Lexi Walker, Cosmin Paduraru, Christina Sorokin, and 1118 others. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *Preprint*, arXiv:2403.05530.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: a large-scale dataset for fact extraction and VERification. In *NAACL-HLT*.

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.

Haoran Wang, Aman Rangapur, Xiongxiao Xu, Yueqing Liang, Haroon Gharwi, Carl Yang, and Kai Shu. 2025. Piecing it all together: Verifying multi-hop multimodal claims. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 7453–7469, Abu Dhabi, UAE. Association for Computational Linguistics.

Lianwei Wu, Yuan Rao, Ling Sun, and Wangbo He. 2021. Evidence inference networks for interpretable claim verification. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(16):14058–14066.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2369–2380, Brussels, Belgium. Association for Computational Linguistics.

Barry Menglong Yao, Aditya Shah, Lichao Sun, Jin-Hee Cho, and Lifu Huang. 2023. End-to-end multimodal fact-checking and explanation generation: A challenging dataset and models. *Preprint*, arXiv:2205.12487.

Xin Yuan, Zhe Lin, Jason Kuen, Jianming Zhang, Yilin Wang, Michael Maire, Ajinkya Kale, and Baldo Faieta. 2021. Multimodal contrastive training for visual representation learning. *Preprint*, arXiv:2104.12836.

Delong Zeng, Yuexiang Xie, Yaliang Li, and Ying Shen. 2025. Enhancing multimodal retrieval via complementary information extraction and alignment. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 22092–22105, Vienna, Austria. Association for Computational Linguistics.

Hang Zhang, Yeyun Gong, Yelong Shen, Jiancheng Lv, Nan Duan, and Weizhu Chen. 2022a. Adversarial retriever-ranker for dense text retrieval. *Preprint*, arXiv:2110.03611.

Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. 2021. Vinvl: Revisiting visual representations in vision-language models. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5575–5584.

Yanzhao Zhang, Dingkun Long, Guangwei Xu, and Pengjun Xie. 2022b. Hlatr: Enhance multi-stage text retrieval with hybrid list aware transformer reranking. *Preprint*, arXiv:2205.10569.

Yi Zhao, Chaozhuo Li, Jiquan Peng, Xiaohan Fang, Feiran Huang, Senzhang Wang, Xing Xie, and Jibing Gong. 2023. Beyond the overlapping users: Cross-domain recommendation via adaptive anchor link learning. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '23, page 1488–1497, New York, NY, USA. Association for Computing Machinery.

Liwen Zheng, Chaozhuo Li, Xi Zhang, Yu-Ming Shang, Feiran Huang, and Haoran Jia. 2024. Evidence retrieval is almost all you need for fact verification. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 9274–9281, Bangkok, Thailand. Association for Computational Linguistics.

Wanjun Zhong, Jingjing Xu, Duyu Tang, Zenan Xu, Nan Duan, Ming Zhou, Jiahai Wang, and Jian Yin. 2020. Reasoning over semantic-level graph for fact checking. In *Proceedings of the 58th Annual Meeting of*

the Association for Computational Linguistics, pages 6170–6180, Online. Association for Computational Linguistics.

Jie Zhou, Xu Han, Cheng Yang, Zhiyuan Liu, Lifeng Wang, Changcheng Li, and Maosong Sun. 2019. GEAR: Graph-based evidence aggregating and reasoning for fact verification. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 892–901, Florence, Italy. Association for Computational Linguistics.

Tianshuo Zhou, Sen Mei, Xinze Li, Zhenghao Liu, Chenyan Xiong, Zhiyuan Liu, Yu Gu, and Ge Yu. 2024. MARVEL: Unlocking the multi-modal capability of dense retrieval via visual module plugin. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14608–14624, Bangkok, Thailand. Association for Computational Linguistics.

A Algorithm

Algorithm 1 Dynamic Adaptive Contrastive Learning Training

Input: Claim set C, evidence corpus E, precomputed fixed negative sample pool, initial model M_0
Initialize: $\bar{\Delta} \leftarrow 0$, $\text{Acc}_{\text{val}} \leftarrow 0$, dynamic negative pool $\mathcal{P}_{\text{dyn}} \leftarrow \emptyset$
for $t = 1$ to T **do**
  $B \leftarrow$ sample a batch from training set
  $\mathbf{h}_c, \mathbf{h}_{e^+}, \mathbf{h}_{\mathcal{N}} \leftarrow M_t(B)$ # compute embeddings
  $\Delta_t \leftarrow$ compute margin following Eq. (1)
  $\bar{\Delta}_t \leftarrow 0.9\bar{\Delta}_{t-1} + 0.1\Delta_t$
  if $t \bmod U_{\text{eval}} = 0$ **then**
    $\text{Acc}_{\text{val}} \leftarrow$ retrieval accuracy on validation set
  end if
  $\Delta_{\text{mid}} \leftarrow \Delta_{\min} + (\Delta_{\max} - \Delta_{\min}) \cdot \text{Acc}_{\text{val}}$
  $p_{\text{dyn}} \leftarrow \sigma((\bar{\Delta}_t - \Delta_{\text{mid}})/\tau_s)$, $\beta \leftarrow 1 - p_{\text{dyn}}$
  Sample negatives: $K \cdot (1 - p_{\text{dyn}})$ from fixed pool, $K \cdot p_{\text{dyn}}$ from $\mathcal{P}_{\text{dyn}}$
  Compute $\mathcal{L}_{\text{full}}$ and $\mathcal{L}_{\text{unit}}$ (with current β)
  $\mathcal{L} \leftarrow \mathcal{L}_{\text{unit}} + \mathcal{L}_{\text{struct}}$
  Backpropagate and update $M_t \rightarrow M_{t+1}$
  if $t \bmod U_{\text{hard}} = 0$ **then**
    Mine hard negatives using M_{t+1}, update $\mathcal{P}_{\text{dyn}}$
  end if
end for
Output: Trained retrieval model M_T

Algorithm 1 describes the dynamic adaptive contrastive learning process. For a given claim c and evidence corpus E, the model maintains both a fixed negative sample pool (precomputed via random and TF-IDF sampling) and a dynamic hard negative pool $\mathcal{P}_{\text{dyn}}$ updated during training. At each training step t, the model computes embeddings for the claim, positive evidence, and sampled negative evidence. The margin Δ_t between positive and negative similarities is calculated following Equation (7), and its exponential moving average $\bar{\Delta}_t$ is updated to reflect the model's current discriminative capability. The validation retrieval accuracy Acc_{val} is periodically evaluated to adaptively calibrate the threshold Δ_{mid}. The dynamic negative sampling ratio p_{dyn} and the structure loss weight β are then derived via a sigmoid mapping: $p_{\text{dyn}} = \sigma((\bar{\Delta}_t - \Delta_{\text{mid}})/\tau_s)$ and $\beta = 1 - p_{\text{dyn}}$. For each claim, $K \cdot (1 - p_{\text{dyn}})$ negatives are drawn from the fixed pool while $K \cdot p_{\text{dyn}}$ are drawn from the dynamic pool $\mathcal{P}_{\text{dyn}}$. The total loss $\mathcal{L} = \mathcal{L}_{\text{full}} + \mathcal{L}_{\text{unit}}$, where $\mathcal{L}_{\text{unit}} = \beta \cdot \mathcal{L}_{\text{sent}} + (1 - \beta) \cdot \mathcal{L}_{\text{struct}}$, is used to update the model via backpropagation. Additionally, every U_{hard} steps, the current model mines new hard negatives to refresh $\mathcal{P}_{\text{dyn}}$. As training progresses, higher margin and validation accuracy lead to increased p_{dyn} and decreased β, enabling the model to gradually shift from easy, fine-grained features to hard, global event-level matching.

B Dataset Statistics

The data statistics of all the datasets used in the experiment is shown in Table 4.

The **FEVER** dataset is a large-scale benchmark dataset for text claim verification, containing 185,445 statements generated by rewriting sentences from Wikipedia. Each claim is labeled as SUPPORTED, REFUTED, or NEI. Among them, 16.82% of the statements require the combination of evidence from multiple sentences, and 12.15% of the statements need to obtain evidence from multiple Wikipedia pages.

The **FEVEROUS** dataset is a large-scale benchmark for fact verification that focuses on mixed non-structured and structured evidence. It contains 87,026 claims generated based on sentences and tables from Wikipedia, each of which is labeled with one of three tags: support, refute, or insufficient information(NEI). Each claim is also associated with specific evidence sentences and table cells. In the dataset, 41% of the statements only contain sentence evidence, 32% only contain table evidence, and 27% require the use of both sentence and table evidence. On average, each statement requires 1.4 sentences and 3.3 cells as evidence.

The **MMCV** dataset is a large-scale benchmark dataset for multi-hop and multi-modal claim verification task, containing 15,569 claims that require reasoning by combining various pieces of evidence such as text, images, and tables. This dataset supports reasoning depths ranging from 1 to 4 hops, with the majority consisting of 1-hop (5,884 claims) and 2-hop (8,485 claims) statements, while also including high-complexity samples with 3 hops (804 claims) and 4 hops (396 claims).

he **Mocheg** dataset is a large-scale benchmark dataset for end-to-end multimodal claim verification. It contains 15,601 real claims, each of which is labeled with a truthfulness tag (support, refute, or

		FEVER	FEVEROUS	MMCV	Mocheg
Annotation	Supported	86,701	49,115	7,361	5,144
	Refuted	36,441	33,669	8,208	5,855
	NEI	42,305	4,242	-	4,602
Modality	Text	-	-	11,815	33,880
	Images	-	-	6,073	12,112
	Tables	-	-	8,962	-

Table 4: This table shows the dataset statistics.

Parameter name	Parameter value
epoch	40
batch size	32
learning rate	1e-5

Table 5: The hyperparameters involved in the training process.

NEI), along with corresponding textual and image evidence, as well as conclusive explanations for the reasoning process. The dataset covers a total of 33,880 pieces of textual evidence and 12,112 pieces of image evidence, and has constructed a retrieval corpus consisting of 91,822 documents and 122,246 images. Only approximately 30% of the documents and 10% of the images contain evidence related to the claims, significantly increasing the difficulty and authenticity of evidence retrieval.

C Implementation Details

In the data preprocessing stage, 150 pages per claim are retrieved by TFIDF and BM25 separately and merged together by ensemble reranking(Dwork et al., 2001) . Furthermore, we selected 30 candidate evidence from each selected documents.

Furthermore, for the bidirectional encoder RoBERTa, the proposed dynamic adaptive contrastive learning is adopted for training optimization. For the RoBERTa, which serves as the cross-encoder, the proposed dynamic negative sample mining method is adopted to provide dynamic negative samples for its training, and the training is conducted using the standard SFT method.

The hyperparameters used in the training are as shown in Table 5.

D Preliminary Experiment

We evaluate the existing general multimodal retrieval methods on the MMCV dataset. The results shows that the performance of the mainstream methods has declined to varying degrees on the multimodal fact checking dataset compared to the general multimodal retrieval dataset, as shown in Table 6. This further supports the viewpoint we present in the second section: The general multimodal retrieval methods cannot be directly applied to evidence-based fact checking retrieval.

Method	WebQA			MMCV		
	MRR@20	NDCG@20	Recall@20	MRR@20	NDCG@20	Recall@20
BM25	22.80	25.41	46.27	21.45	24.24	45.44
VinVL-DPR	38.74	37.79	53.89	34.74	35.66	50.12
CLIP-DPR	49.34	49.11	69.84	39.50	36.23	55.37
UniVL-DPR	62.96	61.22	80.37	41.24	40.27	55.24
T5-ANCE	64.40	64.02	83.81	46.34	47.25	61.35
MARVEL	65.67	65.00	84.32	50.97	48.06	63.46
CIEA	66.41	65.85	85.43	52.67	45.69	67.70

Table 6: Performance comparison on WebQA and MMCV datasets.

E Prompt

The prompt template of MLLM used in the "Event Summary Generation" section is shown in Figure 6.

You are a professional event information extraction expert. Please extract event skeleton information from the given content.

【Extraction Requirements】
1. **Event Summary**: Summarize the complete event in 1-2 sentences, incorporating modality-specific details

2. **Participant Analysis**: Identify all participating entities and their attributes - For images: include visual attributes (appearance, clothing, expressions) - For tables: include data entities (companies, metrics, time periods)

3. **Event Structure**: Use placeholders [ENTITY_TYPE] for specific entities and [ATTRIBUTE_TYPE] for specific attributes

4. **Masked Versions**: Generate both participant-masked and attribute-masked versions

5. **Modality Tags**: Mark which information comes from which modality aspect【Output Format】Please output strictly in the following JSON format:{{ "event_summary": "Complete event summary text incorporating modality details", "participants": [{{ "entity": "Entity name", "type": "Entity type (PERSON/ORG/LOCATION/OBJECT/etc.)", "attributes": ["Attribute1", "Attribute2", ...], }}], "event_structure": "Event skeleton with placeholders, e.g.: [PERSON] at [LOCATION] uses [OBJECT] for [ACTION]", "masked_participants": "Event skeleton with participants replaced by [MASK], preserving attributes", "masked_attributes": "Event skeleton with attributes replaced by [MASK], preserving participants",}}

【Examples】
【here is examples】

Figure 6: The used prompt template