Leveraging Embedding Screening for Multimodal Multi-Hop Claims Verification

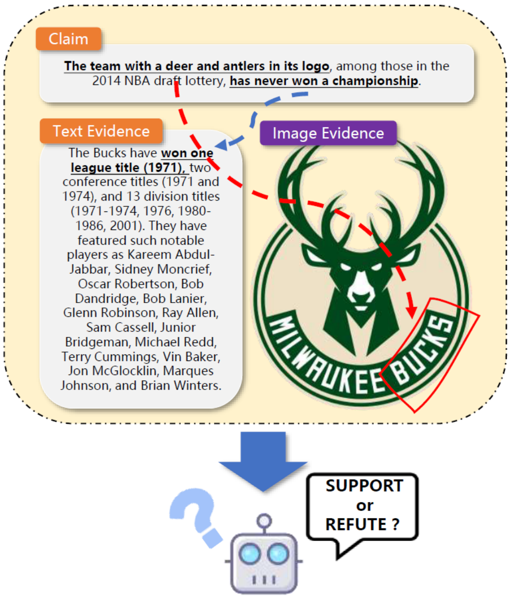
Anonymous ACL submission

Abstract

With the rapid development of generative AI and the explosive growth of Internet, a large amount of multimodal misinformation has been spreading wantonly. Zero-shot claim verification is crucial for combating this issue. Checking a claim requires multi-hop reasoning across evidence with multiple modalities. Consequently, we design a framework called ES4CV, which utilizes **E**mbedding **S**creening **for** multimodal multi-hop **C**laim **V**erification. It consists of two modules: one for zero-shot evidence screening and another for zero-shot claims verification. Within the evidence screening module, we employ a General Multimodal Embedder(GME) to project both multimodal evidence and claims into a unified semantic space, where evidence is screened based on similarity. In the zero-shot claim verification module, the filtered evidence and claims are ultimately fed into a Vision Language Model (VLM) for final judgment. We conduct extensive comparative and ablation experiments on the recently released multimodal multi-hop dataset MMCV to demonstrate our method's effectiveness and superiority.

1 Introduction

With the rapid emergence of new-generation generative artificial intelligence (GAI) represented by large language models (LLMs) and vision-language models (VLMs)(Huang et al., 2025; Zhang et al., 2024), coupled with the accelerated development of social media, massive amounts of GAI-generated synthetic content are flooding the internet, distorting public perception. This phenomenon has been particularly exacerbated in recent years by advanced diffusion models, exemplified by DALL-E(Ramesh et al., 2021) and Stable Diffusion(Rombach et al., 2022), which significantly amplify the scale and realism of AI-generated misinformation.Claim verification is an essential approach in combating misinformation, verifying the authenticity of claim often requires integrating multimodal evidence from diverse sources(Yang et al., 2018), demanding detection systems with multimodal and multi-hop information processing capabilities. Meanwhile, given the complexity and time-sensitive nature of information on contemporary social media platforms(Tasnim, 2020), while domain-specific models may demonstrate strong performance in claim verification within their trained domains, they often fall short when dealing with real-world



The team with a deer and antlers in its logo, among those in the 2014 NBA draft lottery, **has never won a championship**.

The Bucks have **won one league title (1971),** two conference titles (1971 and 1974), and 13 division titles (1971-1974, 1976, 1980-1986, 2001). They have featured such notable players as Kareem Abdul-Jabbar, Sidney Moncrief, Oscar Robertson, Bob Dandridge, Bob Lanier, Glenn Robinson, Ray Allen, Sam Cassell, Junior Bridgeman, Michael Redd, Terry Cummings, Vin Baker, Jon McGlocklin, Marques Johnson, and Brian Winters.

Figure 1: An instance of redundant information in multimodal evidence affecting a model's judgment is verifying the claim: "The team with a deer and antlers in its logo, among those in the 2014 NBA draft lottery, has never won a championship." Actually, just providing the Milwaukee Bucks' logo suffices, as the model can judge based on its prior knowledge. Excessive text may confuse the model, even causing hallucinations and impacting the final judgment.

claims and evidence that have complex compositions.Therefore, zero-shot capability in misinformation detection has become a critical evaluation criterion for detection systems.

To achieve zero-shot multimodal multi-hop claim verification, LLMs and VLMs are often used for their rich general-domain prior knowledge(Liu et al., 2024a).However, the multimodal evidence implies more noise information.As shown in Figure 1, to verify the claim "The team with a deer and antlers in its logo, among those in the 2014 NBA draft lottery, has never won a championship", it is only necessary to provide the VLM with the image evidence of the Milwaukee Bucks' logo. The model can then make a judgment based on this image and its prior knowledge. In the textual evidence, only the phrase "won one league title (1971)" is useful for the model's judgment. The remaining information serves as irrelevant noise, which can induce hallucinations and mislead the model's judgment.

To address this issue, we propose ES4CV, a Embedding Screening based framework for multimodal multi-hop Claims Verification. This framework employs general-domain embedding models to embed claims and multimodal evidence. By filtering the evidence based on similarity, it retains only the evidence highly relevant to the claim. This process effectively cleans the noise in the evidence, thereby improving the model's judgment accuracy. In summary, our contributions are as follows:

- We propose an Embedded Screening based multimodal Multi-hop Claims Verification framework called ES4CV.We uniformly embed the information to be detected and multimodal evidence, and screen the multimodal evidence based on the similarity of these embeddings to reduce the impact of information noise when the model makes judgments.

- During the embedding process,we introduced a General Multimodal Embedder(GME) based on VLM construction, which leveraged the powerful semantic and image understanding capabilities of VLM for embedding. It can map text and images to the same semantic space under zero-shot conditions.

- We demonstrated the effectiveness of our method through comprehensive comparative experiments and ablation experiments, and analyzed in detail the reasons for the hallucinations that occurred when the model made judgments on claims verification.

2 Related Work

In this section, we introduce relevant work on multimodal multi-hop claims verification, including related datasets and research progress on methods. In 2.1, we present the newly proposed MMCV dataset for multimodal multi-hop claims verification. In 2.2 and 2.3, we'll respectively cover feature extraction based and VLM based claims verification methods, along with their inapplicability to real-world claims verification.

2.1 MMCV Dataset

In recent years, numerous scholars have constructed challenging datasets in order to assess model capabilities in multimodal multi-hop misinformation detection. Several teams have developed multimodal claim verification datasets, including FakeNewsNet(Shu et al., 2019), COSMOS(Aneja et al., 2021)and Mocheg(Yao et al., 2023). However, these datasets lack multi-hop reasoning components, thus failing to adequately evaluate models" inference capabilities across sequential information chains.Other research teams have developed multi-hop reasoning datasets, such as QAngaroo(Welbl et al., 2018), ComplexWebQuestion(Talmor and Berant, 2018), and HoVer(Jiang et al., 2020). However, these frameworks predominantly focus on unimodal contexts, failing to account for multimodal contextual information and consequently limiting their evaluation to single-modality reasoning capabilities. In contrast, The MMCV dataset(Wang et al., 2025) was proposed by integrating considerations of both multimodality and multi-hop, enabling a comprehensive assessment of multi-hop reasoning capabilities within a multimodal context.

The MMCV dataset comprises 15k multihop claims paired with multimodal evidence for SUPPORT/REFUTE. It assesses models' ability to combine multimodal evidence for multihop reasoning. The dataset distribution is shown in the Table 1, where n-hop means each claim has n pieces of evidence proving partial truth, requiring n-hop reasoning.

When evaluating VLM on it, we found that multimodal evidence often brings extra noise, causing hallucinations and affecting judgment. Thus, we proposed the ES4CV framework to filter evidence,

Data	1-hop	2-hop	3-hop	4-hop
# Claims	5,884	8,485	804	396
Ave. # Tokens in Claim	21.7	25.32	25.44	26.17
Max. # Tokens in Claim	48	58	51	63
# Text Evidence	2,590	7,323	1,142	760
# Image Evidence	1,979	2,948	634	512
# Table Evidence	1,315	6,699	636	312
# SUPPORT Labels	2,824	4,030	349	158
# REFUTE Labels	3,060	4,455	455	238

Table 1: Data Distribution of MMCV.

retaining high - quality data for model judgment with prior knowledge. To our knowledge, no other team has tested methods on this dataset.

2.2 Claims Verification Method based on feature extraction

Before the introduction of LLMs and VLMs, most research on claim verification focused on using neural networks to extract potential information from multimodal evidence and represent it comprehensively to assist models in determining the authenticity of claims(Liu et al., 2023b; Chen et al., 2023; Khattar et al., 2019). For instance, Safe(Zhou et al., 2020) and BTIC(Zhang et al., 2021) enhanced the representation of multimodal evidence by setting appropriate loss functions. Even after the advent of LLMs and VLMs, this approach has continued to attract numerous researchers due to its excellent performance. Currently, the mainstream method involves aligning the features of entities in claims and evidence to determine the authenticity of claims. CAFE (Chen et al., 2022) calculates the ambiguity between different modal elements using KL divergence, while FND-CLIP (Zhou et al., 2023) achieves outstanding results through element-level semantic detection. Some studies have gone further by leveraging the connections between entities to achieve more precise judgments on the claims verification(Ma et al., 2024). However, the prerequisite for these methods to achieve high accuracy is that they have been comprehensively trained on datasets, and they cannot perform well in a zero-shot setting. Additionally, these methods have poor interpretability, as the models mainly make judgments based on the consistency of features between evidence and claims rather than true reasoning at the level of factual consistency between evidence and claims.

2.3 Claims Verification Method based on VLM

Visual Language Model (VLM) has demonstrated outstanding capabilities in various downstream tasks. Capitalizing on the rich prior knowledge embedded in LLMs and VLMs, research teams have begun exploring Verification frameworks utilizing these large-scale parameterized models(Liu et al., 2024a). Some studies have pointed out the significant potential of multimodal language models in claims verification(Liu et al., 2024b), and some studies have utilized the semantic understanding ability and prior knowledge of VLM to extract more features from the text to be detected(Zheng et al., 2025), in order to assist the classifier in making the final authenticity judgment. However, these studies merely directly make judgments on the claim to be detected, without involving the use of the provided evidence to enable the model to conduct multimodal multi-hop reasoning to reach the final result. In contrast, our method directly employs VLM for the final judgment, effectively leveraging the powerful reasoning ability of VLM to summarize the multimodal evidence, and reducing the influence of noise on the model through evidence screening, thereby improving the accuracy of the model's prediction.

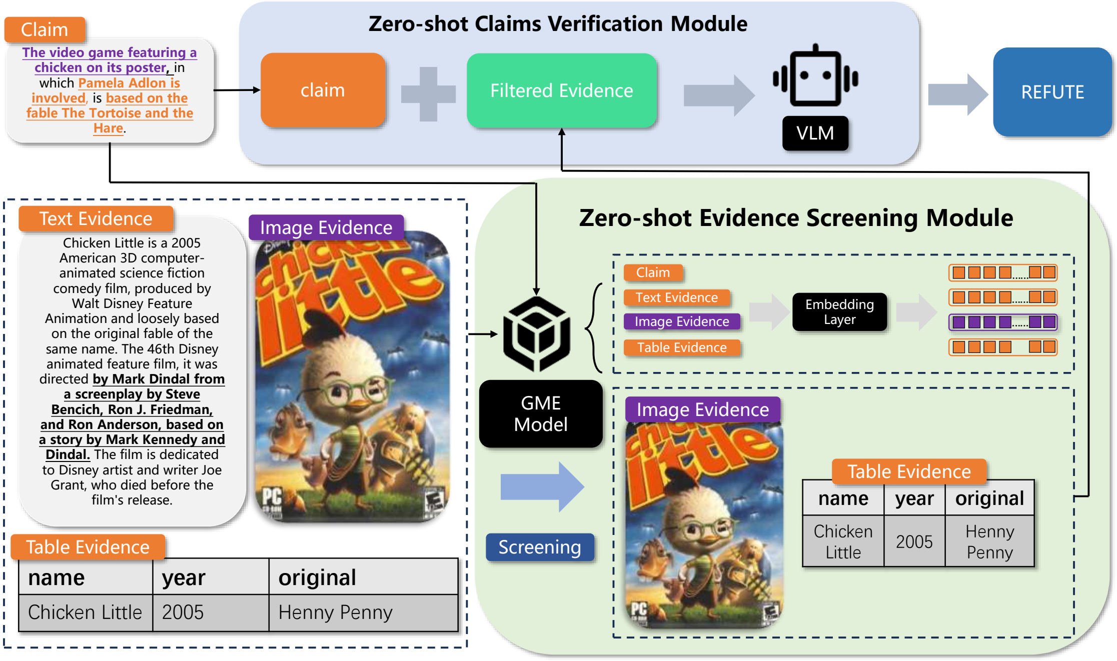
3 Method

In this section, we introduce the Embedded Screening based multimodal multi-hop Claims Verification framework(ES4CV) that we propose. Firstly, we provide a simple definition of the task (3.1) and give an overview of our method (3.2). Then, we elaborate on how our Zero-shot Evidence Screening Module operates (3.3), and how we construct the Zero-shot Claim Verification Module(3.4).

3.1 Task Definition

The problem of multimodal multi-hop claims verification is a typical multimodal binary classification problem. Given a set of cross-modal sample pairs consisting of claims to be detected and a collection of multimodal evidence for reference, the goal of the task is to determine the claim to be detected based on the evidence in the evidence set (which may consist of 1 to 3 pieces), and to provide the final detection result (SUPPORT or REFUTE). Our focus is on detecting metaphors in image-text pairs, so the task can be expressed as:

$$Y = F(x_{cn}, E) \quad (1)$$



The video game featuring a chicken on its poster, in which Pamela Adlon is involved, is based on the fable The Tortoise and the Hare.

Chicken Little is a 2005 American 3D computer-animated science fiction comedy film, produced by Walt Disney Feature Animation and loosely based on the original fable of the same name. The 46th Disney animated feature film, it was directed **by Mark Dindal from a screenplay by Steve Bencich, Ron J. Friedman, and Ron Anderson, based on a story by Mark Kennedy and Dindal.** The film is dedicated to Disney artist and writer Joe Grant, who died before the film's release.

name	year	original
Chicken Little	2005	Henny Penny

Figure 2: An overview of ES4CV.The ES4CV framwork has two parts: a Zero-shot Evidence Screening Module based on GME Model and a VLM - based Zero-shot Claim Verification Module.

where x_{cn} represents the claim to be detected, and the subscript n indicates that this piece of information requires n pieces of evidence for its judgment, meaning that the model needs n steps of thinking to reach the final judgment. E is the set of evidence required to detect this piece of information, and it is given by $E = pte_1, pte_2, pi_1, pi_2, pta_1, pta_2, ..., pte_m, pi_n, pta_k,$ where p_{te_m} is the mth piece of text evidence in the evidence set,p_{i_n} is the nth piece of image evidence, and p_{ta_k} is the kth piece of table evidence. For all three types of evidence, $1 \leq m + n + k \leq 4$ exists. Y represents the classification result of the detected claim, and there are two results: SUPPORT and REFUTE. The focus of our work is to optimize method F to make the model's prediction results as close as possible to the real results.

3.2 Method Overview

Our method overview is shown in Figure 2. ES4CV consists of two parts:the Zero-shot Evidence Screening Module based on GME Model and the Zero-shot Claims Verification Module based on VLM. In the Zero-shot Evidence Screening Module, the claim to be detected and multimodal evidence are sent to the GME model for unified feature encoding. This obtains separate encodings for the claim and evidence in a unified semantic space. Then, cosine similarity between claim and evidence embeddings is calculated. Based on this, evidence with much noise info is excluded.The screened evidence set and claim are sent to the Zero-shot Calim Verification Module for final judgment. In the Zero - shot Claim Verification module, the claim and screened evidence are sent to the VLM. It uses the VLM's rich prior knowledge and reference multimodal evidence to get the final result.

3.3 Zero-shot Evidence Screening Module

To reduce the impact of noise on the model's judgment ability, we construct a Zero-shot Evidence Screening Module based on the multimodal retrieval model General Multimodal Embedder (GME). Through the multimodal embedding layer of GME, we uniformly embed multimodal evidence and claims, and calculate their cosine similarity. Then, we screen the evidence based on the similarity, retaining those with a higher similarity to the claim to be detected (i.e., evidence carrying information highly relevant to the claim and with less noise), and filtering out those with a lower similarity (i.e., evidence carrying more noise and having a lower relevance to the claim), thereby reducing the impact of noise on the model to a certain extent.

3.3.1 GME model for multimodal evidence alignment

The General Multimodal Embedder (GME)(Zhang et al., 2025) is an instruction-based embedding framework, built on the backbone of multimodal large language models (MLLMs), and constructed on the powerful Qwen2-VL series of vision-language models (VLMs). It supports cross-modal retrieval under a unified paradigm, including text, images, visual documents, and fused modalities (i.e., image-text composites). By using the GME model to uniformly embed multimodal evidence sets and claims, it can leverage the general-domain semantic understanding capabilities of its backbone VLM to map these multimodal data into a unified semantic space under zero-shot conditions, laying a solid foundation for various downstream tasks.

We innovatively applied the GME model to aligning semantic features of claims and multi-modal evidence. Leveraging its prior knowledge and semantic understanding, we achieve zero-shot multimodal feature embedding.The embedding formula of GME is as follows:

$$\vec{x_c}, \vec{p_{te}}, \vec{p_i}, \vec{p_{ta}} \ = \ GME(x_c, p_{te}, p_i, p_{ta}) \qquad (2)$$

where $\vec{x_c}$, $\vec{p_{te}}$, $\vec{p_i}$, and $\vec{p_{ta}}$ are the feature vectors obtained after embedding the claim, textual evidence, image evidence, and tabular evidence respectively.

3.3.2 Multimodal Information Noise Filtering Based on Similarity Screening

After obtaining the embeddings corresponding to the claim and the multimodal evidence set through GME, we perform filtering based on the cosine similarity. The formula for cosine similarity is as follows:

$$cosine\_similarity \ = \ \frac{\sum_{i=1}^{d} a_i b_i}{\sqrt{\sum_{i=1}^{d} a_i^2}\sqrt{\sum_{i=1}^{d} b_i^2}} \qquad (3)$$

where vector a and $vectorb$ are both feature vectors of dimension d. The higher the cosine similarity between the two vectors, the closer the distance and direction of the two vectors in the semantic space are, which means the semantic similarity of the two vectors is higher. In the context of evidence screening, the evidence vector with a higher similarity to the claim represents that this piece of evidence carries more evidence information related to the claim.

In order to improve the efficiency of vector calculations, we first perform normalization on the obtained vectors: For a vector a, the vector $a_{normalized}$ obtained by normalizing its L2 norm can be expressed as:

$$a_{normalized} = \frac{a}{\|a\|_2} \qquad (4)$$

where $\|a\|_2$ is the L2 norm of the vector a, and the calculation formula is:

$$\|a\|_2 = \sqrt{\sum_{i=1}^{d} {a_i}^2} \qquad (5)$$

Finally, the dot product is calculated for the two normalized vectors to determine the cosine similarity between the two vectors:

$$cosine\_similarity = \sum_{i=1}^{d} a_i b_i \qquad (6)$$

By plugging $\vec{x_c}$ and the embedded evidence set $\vec{E}$ into the formula, we can obtain the similarity set S_1 between the claim and the evidence from each modality in the evidence set E:

$$\begin{aligned} &\{s_{cte1}, s_{cte2}, ..., s_{ctem}, s_{ci1}, s_{ci2}, ..., s_{cin}, \\ &s_{cta1}, s_{cta2}, ..., s_{ctak}\} \\ &\ = \ S_1 \\ &\ = \ cosine\_similarity(\vec{x_c}, \vec{E}) \end{aligned} \qquad (7)$$

where s_{ctem}, s_{cin}, and s_{ctak} respectively represent the similarity of the corresponding textual evidence, image evidence, and table evidence in the claim and evidence sets. The similarities smaller than z (where z is an adjustable hyperparameter) in S_1 are filtered out. The resulting filtered similarity set S_2 is obtained. Then, based on the mapping relationship between the similarity and the evidence, the corresponding evidence in the evidence set is filtered out to obtain the filtered evidence set E_s. Finally, set E_s would be sent to the Zero-shot Claim Verification Module to make judgments, and the final output result is obtained.

3.4 Zero-shot Claim Verification Module

In order to achieve multimodal multi-hop claim verification under the zero-shot condition, we use VLM to make judgments on the claims to be detected. Similar to LLM, during the training process, VLM usually uses a large amount of image-text pairs of data for learning to discover the correlations and mapping relationships between vision and language, thereby achieving accurate descriptions of images and visual understanding of text. The rich training data enables VLM to possess a large amount of prior knowledge without fine-tuning, and even without providing an evidence set, VLM can already achieve a relatively high accuracy rate for claim verification. Therefore, we chose it as the final judgment model to achieve false information detection under the zero-shot condition, as shown in the following formula:

$$Y = VLM(x_{cn}, E) \tag{8}$$

In the zero - shot claim verification module, we send the filtered evidence set E and the claim to be checked to the VLM. The evidence filtered by the Zero-shot Evidence Screening Module is highly relevant to the claim and has less noise information. This allows the VLM to use the information more efficiently to assist in claim verification, thereby improving the accuracy of zero - shot claim verification.

4 Experiments and Results

4.1 Setup

4.1.1 Baselines

Since there are no other models available for multimodal multi-hop misinformation detection at present, we adopted the method of directly using VLM for judgment as our baseline for comparison. We selected several state-of-the-art VLMs: **GPT-4o**(OpenAI et al., 2024) and **Gemini 1.5 Flash**(Team et al., 2025). Additionally, we included **LLaVA-V1.5-7B**(Liu et al., 2023a) as a representative of open-source models in the experiment. The temperature of all VLMs was set to 0.0, and the maximum token value was set to 5000.

4.1.2 Details

Retrieval modes We set up two different Retrieval modes for comparison, namely the open-book mode, the closed-book mode. In the closed-book mode, the model is required to make a judgment on the claim based on its prior knowledge without being provided with any evidence.In the open-book mode, all the golden evidence related to the claim to be detected is submitted to the model. We categorize our ES4CV method as open - book because in its workflow, all evidence for claim verification is provided to the framework. After filtering, the evidence with less noise is given to the VLM. So, it's logical for us to classify it as open - book and use the open - book approach as a baseline for comparison.

Evidence Provided	**F1 of LLaVa**			
	1-hop	2-hop	3-hop	4-hop
Closed_book	63.57	63.87	66.76	64.64
Text-evidence Only	60.53	62.38	64.46	65.32
Image-evidence Only	67.72	65.57	67.72	66.53
Table-evidence Only	58.27	61.76	63.75	63.25
Open_book	57.21	61.50	63.76	66.42

Table 2: The table shows the experimental results of our pre-experiment. We adopted the LLaVa model as the base model and respectively attempted to provide only text evidence, image evidence or table evidence to the model to observe which modal the noise in the evidence mainly came from.

Prompt Since different prompt enhancement methods have different impacts on VLMs with different parameters, we did not use any prompt enhancement methods (such as chain-of-thought, self-questioning, etc.) in our experiments. Instead, we merely designed three sets of pure prompt templates to instruct the VLM to make judgments on the claim under conditions of providing all evidence, providing no evidence, and providing partial evidence. The complete prompt templates will be provided in the appendix.Among them, the open book mode is the baseline method we have chosen.

Evidence screening threshold In the previous text, we mentioned that after obtaining the embedding similarity between the evidence and the claim, we will select an adjustable hyperparameter m as the similarity threshold to screen the evidence. After the ablation study (the content here will be detailed in Section 4.2), we decided to set the screening threshold for textual evidence and tabular evidence at 0.7, and the threshold for image evidence at 0.5.

Evaluation metrics We adopted precision, recall and F1 score as the evaluation criteria for our experiment.

4.2 Pre-Experiment

To explore which modality the noise in the evidence comes from, we conduct a pre-experiment using LLaVa, which generates the most hallucinations beacause of the least parameters. The experimental results are shown in Table 2. We conduct experiments by providing the model with only text, image, and table evidence respectively, and compared the F1 values in open-book and closed-book modes as the baseline.

From the results, it can be seen that when only image evidence is provided to the model, the F1 value has a significant increase compared to the closed-book mode. In contrast, if only text evidence is provided, the F1 value drops significantly. This phenomenon is even more pronounced when only table evidence is provided. We believe that this result is due to the fact that text evidence often contains a large amount of content irrelevant to the claim to be verified. The same is true for table evidence; verifying a claim often only requires specific cells in the table rather than the entire table. Excessive and irrelevant text brings a very serious hallucination tendency to the model. In contrast, image evidence is different; the information carried by images is often closely related to the claim. This is also consistent with human perception: when you search for a certain piece of information, a large amount of text can make you feel lost, while images can present the information you are looking for more clearly. In conclusion, we believe that the noise that disrupts the model mostly comes from text and tables. Therefore, we set the similarity threshold for text and tables to 0.7, and the screening threshold for image evidence to 0.5.

4.3 Main Results

The results of our comparative experiments are shown in Table 3, and for each model, the better scores in open - book and ES4CV modes are highlighted.

According to the experimental results, we can classify the models into two categories. Models with higher closed-book performance than open-book, like LLaVa and GPT-4o, are likely weaker in summarizing evidence and extracting useful information and more susceptible to noise. Conversely, models with higher open-book performance than closed-book, such as Gemini, are less affected by noise and less prone to hallucinations.

Overall, the experimental results prove that our ES4CV framework is effective on models that more susceptible to noise. Most of the results of the LLaVa model and the GPT-4o model on the four sub-datasets have been significantly improved compared with the open-book experiments used as the baseline method. Only the performance of the LLaVa model on the 2-hop sub-dataset were slightly lower than the baseline methods. One point that needs to be particularly noted is that our method is to screen the evidence that contains more noise, thereby helping the models that will be affected by noise to select high-quality evidence. However, the Gemini model itself achieved better results in the open-book experiment than the closed-book one,and there is less evidence related to each claim in 1-hop and 2-hop. Therefore, the effect of our method on Gemini is not good. We think this is in line with expectations. And as the number of reasoning hops increases, the evidence and the noise it contains also increase. It can also be observed from the results that our method has begun to play a certain positive role for Gemini.

Specifically, in the open-book mode, Gemini 1.5 performed the best, with an average F-1 value of 70.92. However, in the closed-book mode, LLaVA surprisingly performed the best, with an average F-1 value of 66.77. Nevertheless, this does not mean that its prior knowledge is richer than that of the GPT model and the Gemini model. Previous studies(Wang et al., 2025) have pointed out that even when LLaVA's predictions are correct, its reasoning process may still have illusions and be a wrong reasoning process. In addition, GPT - 4o is most affected by ES4CV, with an average metric improvement of 6.68 percentage points. This is likely because GPT - 4o has strong reasoning ability but weaker information selection ability. Thus, when the external framework helps with information filtering, its reasoning ability is better unleashed.

5 Conclusion

The ES4CV framework we propose leverages the extensive prior knowledge and powerful reasoning ability of VLM for zero-shot claims verification. Meanwhile, it builds a Zero-shot Evidence Screening Module centered on the multimodal embedding model GME to filter evidence, thereby retaining evidence highly relevant to the claim and screening out multimodal evidence with more noise. This effectively reduces the impact of model hallucinations on judgment results and improves the accuracy of the judgment.

Additionally, we have comprehensively demonstrated the effectiveness of our method through experiments, and detailedly analyzed the reasons for hallucinations when the model conducts multimodal multi-hop reasoning, providing valuable solutions and research contributions to the field of

			1-hop			2-hop			3-hop			4-hop		
Retrieval Mode	**Model**		P	R	F1	P	R	F1	P	R	F1	P	R	F1
Open_book	GPT-4o	Base	76.95	72.95	71.78	68.03	63.24	60.53	62.67	58.78	56.08	67.75	62.46	61.35
		ES4CV	**80.11**	**76.45**	**76.21**	**75.14**	**67.72**	**66.03**	**75.54**	**68.90**	**68.66**	**75.79**	**66.68**	**66.64**
	Gemini	Base	**79.58**	**79.25**	**79.20**	**72.38**	**71.85**	**71.66**	**66.37**	**65.90**	**65.86**	67.21	**66.86**	**66.97**
		ES4CV	77.93	74.29	73.92	71.86	64.69	62.45	65.77	61.50	60.50	**69.70**	64.56	64.52
	LLaVa	Base	62.86	59.68	57.21	64.17	62.48	**61.50**	65.47	64.64	63.76	66.50	66.76	66.42
		ES4CV	**65.35**	**61.84**	**58.84**	**66.00**	**63.16**	60.66	**70.01**	**68.59**	**66.18**	**68.88**	**69.51**	**67.75**
Closed_book	GPT-4o		76.86	72.94	71.79	67.96	63.30	60.66	62.88	58.89	56.17	67.93	62.39	61.20
	Gemini		75.67	71.44	70.15	69.10	64.19	61.73	66.74	61.10	58.44	63.78	59.90	58.69
	LLaVa		64.18	63.78	63.57	64.06	63.93	63.87	66.78	66.81	66.76	64.64	64.84	64.64

Table 3: This table shows our experimental results. We selected three VLMs, namely GPT-4o, Gemini and LLava, for our experiment. The experiment adopted two different modes: the open-book mode, that is, providing all the evidence to the model; Closed-book mode, that is, no evidence is provided to the model, allowing the model to make judgments based on its own prior knowledge. We categorize our ES4CV method as a sub-mode of open-book mode.

multimodal multi-hop misinformation detection.

Limitations

The ES4CV we proposed filters and screens multimodal evidence through the method of evidence embedding similarity screening. This is a coarse-grained screening approach. For instance, if the similarity between a text evidence and the claim to be detected is less than 0.7, the entire text evidence will be completely filtered out. In fact, if we look at the problem from a fine-grained perspective, we will find that not the entire text evidence is irrelevant to the claim to be detected. As shown in Figure 2, although the longer text evidence was filtered out as a whole after screening, the highlighted short sentence "it was directed by Mark Dindal from a screenplay by Steve Bencich, Ron J. Friedman, and Ron Anderson, based on a story by Mark Kennedy and Dindal." is actually high-quality evidence that is helpful for the model's judgment. Therefore, we believe that it is possible to conduct research on evidence from a fine-grained perspective to achieve more precise screening of multimodal evidence.

Furthermore, although our method has been proven effective in comparison and ablation experiments and has achieved a comprehensive improvement over the baseline method that only uses VLM in open-book experiments, its metrics are still lower than those in closed-book experiments. If this issue is resolved, models can better utilize multimodal contextual evidence for more accurate claim evaluation.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (Grant No. 62266048) under the project entitled "Research on Key Problems in National Language Speech Recognition and Synthesis Based on Deep Learning."

References

Shivangi Aneja, Chris Bregler, and Matthias Nießner. 2021. Cosmos: Catching out-of-context misinformation with self-supervised learning. *Preprint*, arXiv:2101.06278.

Yixuan Chen, Dongsheng Li, Peng Zhang, Jie Sui, Qin Lv, Lu Tun, and Li Shang. 2022. Cross-modal ambiguity learning for multimodal fake news detection. In *Proceedings of the ACM Web Conference 2022*, WWW '22, page 2897–2905, New York, NY, USA. Association for Computing Machinery.

Ziwei Chen, Linmei Hu, Weixin Li, Yingxia Shao, and Liqiang Nie. 2023. Causal intervention and counterfactual reasoning for multi-modal fake news detection. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 627–638, Toronto, Canada. Association for Computational Linguistics.

Baixiang Huang, Canyu Chen, and Kai Shu. 2025. Authorship attribution in the era of llms: Problems, methodologies, and challenges. *Preprint*, arXiv:2408.08946.

Yichen Jiang, Shikha Bordia, Zheng Zhong, Charles Dognin, Maneesh Singh, and Mohit Bansal. 2020. HoVer: A dataset for many-hop fact extraction and claim verification. In *Findings of the Association* *for Computational Linguistics: EMNLP 2020*, pages 3441–3460, Online. Association for Computational Linguistics.

Dhruv Khattar, Jaipal Singh Goud, Manish Gupta, and Vasudeva Varma. 2019. Mvae: Multimodal variational autoencoder for fake news detection. In *The World Wide Web Conference*, WWW '19, page 2915–2921, New York, NY, USA. Association for Computing Machinery.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023a. Visual instruction tuning. *Preprint*, arXiv:2304.08485.

Hui Liu, Wenya Wang, and Haoliang Li. 2023b. Interpretable multimodal misinformation detection with logic reasoning. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 9781–9796, Toronto, Canada. Association for Computational Linguistics.

Hui Liu, Wenya Wang, Haoru Li, and Haoliang Li. 2024a. TELLER: A trustworthy framework for explainable, generalizable and controllable fake news detection. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 15556–15583, Bangkok, Thailand. Association for Computational Linguistics.

Qiang Liu, Xiang Tao, Junfei Wu, Shu Wu, and Liang Wang. 2024b. Can large language models detect rumors on social media? *Preprint*, arXiv:2402.03916.

Zihan Ma, Minnan Luo, Hao Guo, Zhi Zeng, Yiran Hao, and Xiang Zhao. 2024. Event-radar: Event-driven multi-view learning for multimodal fake news detection. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5809–5821, Bangkok, Thailand. Association for Computational Linguistics.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, and 262 others. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. 2021. Zero-shot text-to-image generation. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8821–8831. PMLR.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695.

Kai Shu, Deepak Mahudeswaran, Suhang Wang, Dongwon Lee, and Huan Liu. 2019. Fakenewsnet: A data repository with news content, social context and spatialtemporal information for studying fake news on social media. *Preprint*, arXiv:1809.01286.

Alon Talmor and Jonathan Berant. 2018. The web as a knowledge-base for answering complex questions. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 641–651, New Orleans, Louisiana. Association for Computational Linguistics.

Samia Tasnim. 2020. Impact of rumors and misinformation on covid-19 in social media. *Journal of preventive medicine and public health = Yebang Uihakhoe chi.*

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M. Dai, Anja Hauth, Katie Millican, David Silver, Melvin Johnson, Ioannis Antonoglou, Julian Schrittwieser, Amelia Glaese, Jilin Chen, Emily Pitler, Timothy Lillicrap, Angeliki Lazaridou, and 1332 others. 2025. Gemini: A family of highly capable multimodal models. *Preprint*, arXiv:2312.11805.

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.

Johannes Welbl, Pontus Stenetorp, and Sebastian Riedel. 2018. Constructing datasets for multi-hop reading comprehension across documents. *Transactions of the Association for Computational Linguistics*, 6:287–302.

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. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '23, page 2733–2743. ACM.

Qihui Zhang, Chujie Gao, Dongping Chen, Yue Huang, Yixin Huang, Zhenyang Sun, Shilin Zhang, Weiye Li, Zhengyan Fu, Yao Wan, and Lichao Sun. 2024. LLM-as-a-coauthor: Can mixed human-written and machine-generated text be detected? In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 409–436, Mexico City, Mexico. Association for Computational Linguistics.

Wenjia Zhang, Lin Gui, and Yulan He. 2021. Supervised contrastive learning for multimodal unreliable news detection in covid-19 pandemic. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, CIKM '21, page 3637–3641, New York, NY, USA. Association for Computing Machinery.

Xin Zhang, Yanzhao Zhang, Wen Xie, Mingxin Li, Ziqi Dai, Dingkun Long, Pengjun Xie, Meishan Zhang, Wenjie Li, and Min Zhang. 2025. Gme: Improving universal multimodal retrieval by multimodal llms. *Preprint*, arXiv:2412.16855.

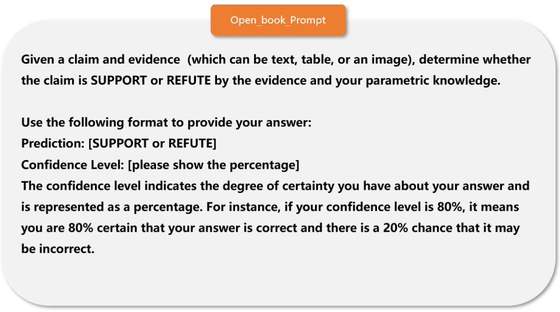
Xiaofan Zheng, Minnan Luo, and Xinghao Wang. 2025. Unveiling fake news with adversarial arguments generated by multimodal large language models. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 7862–7869, Abu Dhabi, UAE. Association for Computational Linguistics.

Xinyi Zhou, Jindi Wu, and Reza Zafarani. 2020. : Similarity-aware multi-modal fake news detection. In *Advances in Knowledge Discovery and Data Mining: 24th Pacific-Asia Conference, PAKDD 2020, Singapore, May 11–14, 2020, Proceedings, Part II*, page 354–367, Berlin, Heidelberg. Springer-Verlag.

Yangming Zhou, Yuzhou Yang, Qichao Ying, Zhenxing Qian, and Xinpeng Zhang. 2023. Multimodal fake news detection via clip-guided learning. In *2023 IEEE International Conference on Multimedia and Expo (ICME)*, pages 2825–2830.

A Example Appendix

A.1 Prompt

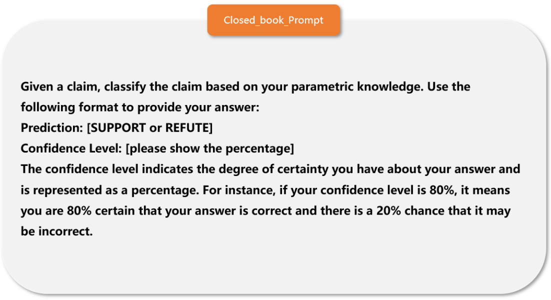


Open_book_Prompt

Given a claim and evidence (which can be text, table, or an image), determine whether the claim is SUPPORT or REFUTE by the evidence and your parametric knowledge.

Use the following format to provide your answer:
Prediction: [SUPPORT or REFUTE]
Confidence Level: [please show the percentage]
The confidence level indicates the degree of certainty you have about your answer and is represented as a percentage. For instance, if your confidence level is 80%, it means you are 80% certain that your answer is correct and there is a 20% chance that it may be incorrect.

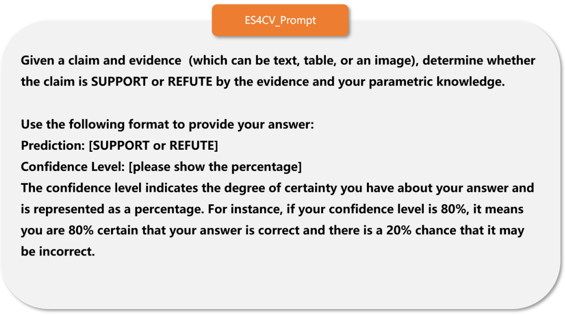
Figure 3: The prompt template we used in the open-book experiment.



Closed_book_Prompt

Given a claim, classify the claim based on your parametric knowledge. Use the following format to provide your answer:
Prediction: [SUPPORT or REFUTE]
Confidence Level: [please show the percentage]
The confidence level indicates the degree of certainty you have about your answer and is represented as a percentage. For instance, if your confidence level is 80%, it means you are 80% certain that your answer is correct and there is a 20% chance that it may be incorrect.

Figure 4: The prompt template we used in the closed-book experiment.



ES4CV_Prompt

Given a claim and evidence (which can be text, table, or an image), determine whether the claim is SUPPORT or REFUTE by the evidence and your parametric knowledge.

Use the following format to provide your answer:
Prediction: [SUPPORT or REFUTE]
Confidence Level: [please show the percentage]
The confidence level indicates the degree of certainty you have about your answer and is represented as a percentage. For instance, if your confidence level is 80%, it means you are 80% certain that your answer is correct and there is a 20% chance that it may be incorrect.

Figure 5: The prompt template we used in the ES4CV experiment.