Blank-filling: Missing Modality-Simulated Network for Robust Multimodal Fact Verification

Anonymous ACL submission

Abstract

Recently, multimodal fact verification tasks aim to assess the truthfulness of multimodal claims by the retrieved evidence through textual and visual content. In contrast, the multimodal information may be incomplete in original posts or missing during the data collection. However, recent missing-modality studies still cannot properly handle the above complex missing situations of claim-evidence input pairs in multimodal fact verification, as they fail to capture complicated relations between claims and evidence. To solve these problems, we propose a novel model named Missing Modality-Simulated Network (MMSN) for more robust and adaptive multimodal fact verification. We design a novel dual-channel soft simulation module to use both cross-modal information and claim-evidence correlations to simulate missing features with a soft-weighted method. Besides, MMSN exploits fine-grained textual key information and designs coarse-grained and fine-grained fusions to fuse multimodal information and capture their interactions exhaustively. The experimental results on three real-world public datasets show the superiority and effectiveness of MMSN for robust multimodal fact verification.

1 Introduction

Fact verification, aiming to assess the truthfulness of claims by the retrieved evidence, has attracted a great amount of attention in research fields (Murayama, 2021; Varnosfaderani et al., 2024; Kanaani, 2024; Zhang et al., 2024; Si et al., 2023; Kim et al., 2023; He et al., 2021). With the rapid development of social platforms, the dissemination of misinformation becomes easier in a multimodal way with textual and visual content. Traditional methods, only leveraging textual information to verify claims, fail to detect fake news and claims with multimodal content. Therefore, multimodal fact verification has become a research hotspot.

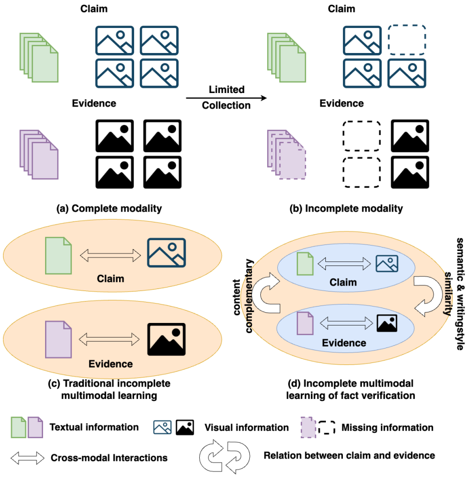


Figure 1: The comparison of the hypothesis of complete modality, the real-world scenario of incomplete modality, and the common framework of traditional methods. Because of the limited collection methods, claims and evidence may miss some information about texts and images in the real world.

Traditional multimodal studies extract textual and visual features and leverage several modality fusion mechanisms to learn multimodal representations and predict the verdict, such as attention-based methods (Mishra et al., 2022; Wang and Peng, 2022; Gao et al., 2022; Zhang et al., 2023) and graph-based methods (Cao et al., 2024; Dhawan et al., 2023; Zhao et al., 2023; Qi et al., 2023).

Recent multimodal fact verification models fail to resolve the real-world data which commonly comes with incomplete information. They are designed based on the hypothesis that both textual and visual modalities are accessible, as Figure 1 (a) shows. However, because of the collection of data which leads to incomplete data (Sun et al., 2023), like Figure 1 (b), textual and visual information may be incomplete in the real world. The aforementioned methods would struggle to make accurate predictions due to the missing patterns. There are some methods of learning high-quality multimodal representations to solve the incompleteness problem (Wen et al., 2023; Sun et al., 2023; Ji et al., 2019; Liu et al., 2024; Yuan et al., 2024; Wang et al., 2023). However, multimodal fact verification entails **cross-modal understandability** as well as **insightful comprehension between claims and evidence** as shown in Figure 1, which makes these methods unsuitable to solve this problem. Specifically, claims and evidence are highly correlated, thus there must be some explicit content overlaps or writing style similarities, which benefits the simulation of missing parts of both claims and evidence. These relations are crucial to multimodal fact verification as well as to simulate missing parts of the evidence, for they can help the model better understand how the evidence supports or refutes the claim (Akhtar et al., 2023; Yao et al., 2023).

To solve the aforementioned problems, we propose a novel model named Missing Modality-Simulated Network (MMSN) for robust multimodal fact verification tasks under missing modality situations. We design a novel Dual-Channel Soft Simulation (DCSS) module to simulate the missing features using multimodal information and claim-evidence relations. It emphasizes insightful comprehension between claims and evidence which is significant and beneficial for multimodal fact verification. Instead of using a hard zero-filling or one-filling way, we utilize a soft weighted filling and simulation method to capture more correlations within and between modalities. Simultaneously, we extract valuable key phrases of textual content as complementary knowledge via a large language model, since fine-grained knowledge enables the model to capture more comprehensive features. Then, we design a Multi-granularity Multimodal Fusion (MMF) module based on attention mechanisms to integrate multimodal and multi-granularity features comprehensively for the prediction of each claim-evidence pair. The DCSS and MMF modules ensure the model learns a robust multimodal representation even the data information is incomplete.

To investigate the performance of our proposed model, we conduct extensive experiments on three commonly used datasets, FACTIFY (Mishra et al., 2022), MOCHEG (Yao et al., 2023), and Fin-Fact (Rangapur et al., 2023). The experimental results show the effectiveness and superiority of our proposed model. We further demonstrate the efficacy of the DCSS and the MMF modules compared with other simulation and fusion methods through several experiments.

Our main contributions are as follows: (1) We propose a novel Missing Modality-Simulated Network for robust multimodal fact verification tasks with incomplete data, enhancing the robustness of the multimodal fact verification model; (2) We propose a novel Dual-Channel Soft Simulation module to comprehensively simulate missing features considering both cross-modal information and claim-evidence correlations with a soft-weighted method, emphasizing the importance of cross-modal correlations and obtaining comprehensive simulated representations; (3) To evaluate the performance of our proposed method, we carry out experiments on three commonly used multimodal datasets. Our model outperforms the comparison methods, which demonstrates the effectiveness and superiority of the proposed model.

2 Methodology

In this section, we present the Missing Modality-Simulated Network (MMSN) in detail for multimodal fact verification with missing visual modality. We begin by providing the problem definition and feature extraction, after which we introduce the overall framework of MMSN. Subsequently, we describe the details of the proposed method.

2.1 Problem definition

Multimodal fact verification aims to verify the given claim with textual and visual contents, using retrieved multimodal evidence from databases, such as Wikipedia and fact-checking websites. Let $\mathcal{P} = \{C_T, C_I, E_T, E_I\}^{|\mathcal{P}|}$ be the corpus of the dataset, where C_T and C_I denote the text and image of the claim, and E_T and E_I denote the text and image of the evidence. Each claim-evidence pair may have complete information or missing text or image information of claim or evidence[1]. Specifically, for each claim-evidence pair p, we manually set 3 flags f_C^I, f_E^T, and f_E^I to indicate whether the text or image of the claim and evidence are missing. If the evidence text is missing, f_E^T is set to 1, otherwise f_E^T is set to 0. Each claim-evidence pair will not simultaneously miss 2 modalities of evidence, that is, $f_E^T \oplus f_E^I = 1$. The target $y \in \mathcal{Y}$. The goal

[1] Here we do not consider the circumstance that both text and image of evidence are missing or claim text is missing, because this kind of data will be regarded as less check-worthy claims and be filtered at the beginning of fact-checking procedure.

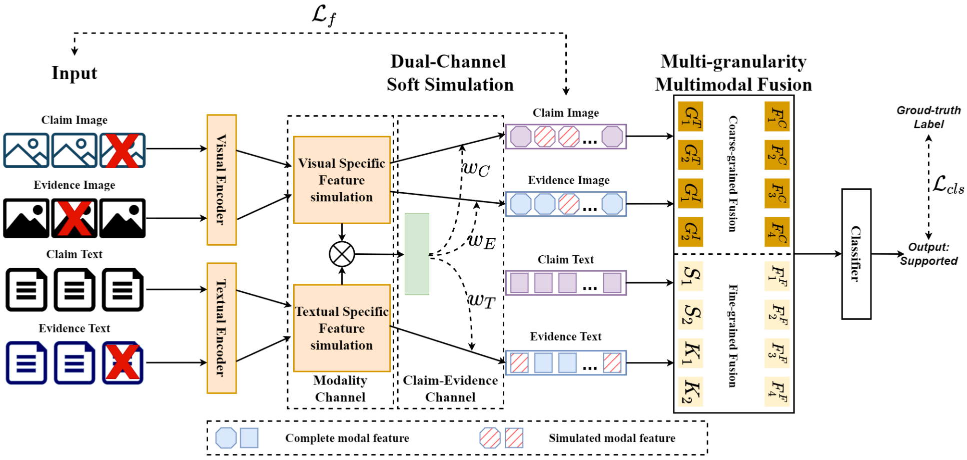


Figure 2: The overall architecture of MMSN. Firstly, given the text and image of the claim and evidence, we leverage pre-trained models (e.g., DeBERTa and SWIN) to extract textual and visual embeddings respectively. Then we enhance the embeddings by extracting key phrases using large language models, to improve the modeling of fine-grained semantics. Based on enhanced textual and visual embeddings, we fill in the missing modality features through the soft simulation module by considering both cross-modal information and claim-evidence correlations. Thirdly, we employ a multi-granularity multimodal fusion module to obtain the multimodal representation and finally make predictions. **G, S, K** denote document-level representations of claim and evidence, sentence-level representations of claim and evidence, and representations of key information, respectively.

is to find a function $F : \mathcal{P} \rightarrow \mathcal{Y}$ that maps the data to the label set and makes predictions.

2.2 Overall architecture of MMSN

Our objective is to learn fine-grained multimodal representations by capturing more granular information in textual and visual content with missing visual modality. To this end, we propose a novel *Missing Modality-Simulated Network* for robust multimodal fact verification. Figure 2 illustrates the overall architecture of MMSN, which mainly consists of the following components: (1) **Feature Extraction and Key Phrase Capture**, (2) **Dual-Channel Soft Simulation**, (3) **Multi-granularity Multimodal Fusion**, and (4) **Classifier**.

2.3 Feature extraction and key phrase capture

First, we extract raw features of both claim and evidence by textual and visual encoders and capture key information of claim texts through key phrase extraction for fine-grained information integration.

Feature extraction Following previous works (Cao et al., 2024), we extract raw sentence-level features of textual content by DeBERTa (He et al., 2021). For the visual modality, following previous works (Cao et al., 2024; Wang and Peng, 2022), we leverage a pre-trained model SWIN (Liu et al., 2021) to get raw visual embeddings. Specifically, we mean-pool the last hidden state of all tokens as the final raw embeddings for both textual and visual embeddings. To integrate the calculation process, we initialize the missing embeddings as 0. t_C, v_C, t_E,and v_E denote the final initialized embeddings of the claim-evidence pair p from textual and visual modality, respectively.

Key phrase capture For fine-grained key phrases, we resort to ChatGPT[2] to extract fine-grained key phrases and obtain fine-grained knowledge. We use the following prompt to extract key phrases:

> *Please return a list containing key phrases that help verify the claim's truthfulness. Let's think step by step. Please return in this form: Key phrases: [key phrases]. Here is the text: [TEXT].*

Then we also utilize DeBERTa to extract raw features and mean-pool the hidden state of all tokens as embeddings of key phrases k_C for claims and evidence, respectively.

2.4 Dual-channel soft simulation

Our objective is to learn comprehensive multimodal representations to verify the truthfulness with missing modalities. These missing contents are crucial to make more accurate predictions and promote the accuracy of detection. Hence, we attempt to utilize

[2] https://openai.com/chatgpt/

a Dual-Channel Soft Simulation (DCSS) to learn and imitate the missing features from available textual and visual features, capturing both intra-modal and inter-modal correlations that simulate the missing representation to the greatest extent.

2.4.1 Modality channel simulation

Considering the real-world scenario, claim text is always available, otherwise it cannot be detected. Hence, we only simulate textual representations of evidence and visual representations of both claims and evidence.

Textual-specific feature simulation For claim-evidence pairs without evidence textual content, we leverage available textual information to extract unique textual features $t_C^{'}$ and $t_E^{'}$ by:

$$t_M^{'} = \sigma(t_M W_1 + b_1), M \in \{C, E\}, \tag{1}$$

where W_1 and b_1 are learnable parameters, and σ denotes the activation function.

Visual-specific feature simulation For visual content simulation, we first utilize available visual representations of both claims and evidence to obtain Visual-specific feature representation $v_C^{'}$ and $v_E^{'}$ by:

$$v_M^{'} = \sigma(v_M W_2 + b_2), M \in \{C, E\}, \tag{2}$$

where W_2 and b_2 are learnable parameters.

Cross-modal correlated representation To capture cross-modal correlations and avoid the disturbance of raw embeddings, we utilize textual and visual unique features to calculate correlated representation for claim-evidence channel simulation Cor_r by:

$$\begin{gathered} t_U = \begin{bmatrix} t_C^{'}, \\ t_E^{'} \end{bmatrix}, v_U = \begin{bmatrix} v_C^{'}, \\ v_E^{'} \end{bmatrix}, \\ Cor_r = \sigma((t_U {v_U}^T) W_3 + b_3), \end{gathered} \tag{3}$$

where t_U and v_U denote textual- and visual-specific representations respectively.

2.4.2 Claim-evidence channel simulation

We leverage the modality-specific features and cross-modal correlated representations to calculate the simulated textual and visual representations. First we calculate the simulated textual representation by:

$$\begin{aligned} \hat{t}_E &= t_E^{'} + w_T Cor_r v_U, \\ t_E^{*} &= t_E + f_E^T \hat{t}_E, \end{aligned} \tag{4}$$

where f_E^T is the flag that demonstrates whether the textual evidence is missing and w_T is the weight calculated by:

$$w_T = softmax(\frac{t_E^{'} {t_U}^T}{\sqrt{d}}), \tag{5}$$

where d denotes the dimension of textual embeddings.

Similarly, we simulate the missing visual representations by:

$$\begin{aligned} \hat{v}_M &= v_M^{'} + w_M {Cor_r}^T t_U, M \in \{C, E\} \\ v_M^{*} &= v_M + f_M^I \hat{v}_M, \end{aligned} \tag{6}$$

where f_M^T is the flag that demonstrates whether the visual representation is missing and w_M is the weight calculated by:

$$w_M = softmax(\frac{v_M^{'} {v_U}^T}{\sqrt{d}}), M \in \{C, E\}. \tag{7}$$

Through the above operations, we leverage cross-modal and claim-evidence features to simulate the missing textual and visual representation and obtain the simulated representations t_C, t_E^*, v_C^*, and v_E^*. To restrain the quality of simulated representations, we utilize the similarity loss $\mathcal{L}_f$ to restrain the quality of simulated representations (see details in section 2.7).

2.5 Multi-granularity multimodal fusion

Through the above procedures, we obtain the available textual representations c_T and e_T, the available visual representations c_I and e_I, the simulated representation of missing modalities $\hat{T}_E$, $\hat{V}_C$, and$\hat{V}_E$, and the representations kp_C and kp_E. Inspired by Zhang et al. (2023), to take full advantage of these features and comprehensively fuse the multimodal representations, we design two kinds of fusion methods to deal with multi-granular features.

We propose a coarse-grained attention fusion module for the coarse-grained data, such as document-level and image-level representations, to capture coarse-grained multimodal interactions. For a claim-evidence pair $\mathcal{C} = \{c_T, c_I^i(\hat{V}_C), e_T(\hat{T}_E), e_I(\hat{V}_E)\}$, the fusion representations $\tilde{m}^{caf}$ are calculated by [3]:

$$\tilde{m}^{caf} = concat(\alpha_1 \Theta_1 c + e, \alpha_2 \Theta_2 e + c), \tag{8}$$

[3] Here we only demonstrate the calculation process of cross-modal representation, actually $\tilde{m}^{caf}$ also contains intra-modal fused representation, and so does $\tilde{m}^{faf}$

where Θ_1 and Θ_2 denote learnable parameters, and $c \in \{c_T, c_I(\hat{V}_C)\}$ and $e \in \{e_T^{,} e_I(\hat{V}_E)\}$. α_1 and α_2 denote the attention scores between claim and evidence:

$$\alpha_1 = softmax(\frac{ec^T}{\sqrt{d}}), \tag{9}$$

$$\alpha_2 = softmax(\frac{ce^T}{\sqrt{d}}). \tag{10}$$

We also propose a fine-grained attention fusion module to capture fine-grained semantics and relations between key phrases and textual representations. Specifically, for key phrases from the claim text, we divide the evidence document into sentences and obtain the sentence representation set e_{sent} [4]. Then we regard each key information representation as a Query to calculate the attention score and obtain fine-grained multimodal representations $\tilde{m}^{faf}$ through a sliding window whose length is l:

$$\tilde{m}^{faf}_{i,j} = mean(kp^i_C + \alpha_3 \Theta_3 e^{l_j}_{sent}), \tag{11}$$

where Θ_3 is a learnable parameter and α_3 denotes the attention scores between textual content and key information, calculated the same as α_1 and α_2, and l_j denotes the j-th time that the window slides.

Then we mean-pool the coarse-grained multimodal features $\tilde{m}^{caf}$ and the fine-grained multimodal features $\tilde{m}^{faf}$ and concatenate them to get the comprehensive multimodal representations $\tilde{m}$ for label prediction.

2.6 Classifier

To predict the label of the given claim-evidence pair, we use the multimodal representation $\tilde{m}$ as input to the category classifier, which consists of a 2-layered fully connected network. The prediction process is carried out as follows:

$$\hat{y} = softmax(W^1 \sigma(W^0 \tilde{m})), \tag{12}$$

where W^0 and W^1 are learnable parameters and $\hat{y}$ is the predicted label.

2.7 Training loss

During the training stage, we calculate two loss functions to restrain the quality of simulated representations and classification accuracy.

First, we design a similarity loss $\mathcal{L}_f$ to calculate the similarity of simulated representations and their origin representations:

$$\begin{aligned}\mathcal{L}_f = &\sum_{i=1}^{|C|}(D(c^i_I, \hat{V}^i_C) f^I_C + D(e^i_T, \hat{T}^I_E) f^T_E \\ &+ D(e^i_I, \hat{V}^i_E) f^I_E),\end{aligned} \tag{13}$$

where f^I_C, f^T_E, and f^I_E are flags to mark whether textual or visual modality is missing (as mentioned in section 2.1), $|C|$ denotes the length of the dataset, and D denotes the distance function. Here we use the cosine similarity as the distance function.

Then we utilize cross-entropy loss $\mathcal{L}_{ce}$ as the classification loss to restrain the model from obtaining a higher performance:

$$\mathcal{L}_{cls} = -\sum_{i=1}^{|C|} y_i log(\hat{y}_i), \tag{14}$$

where $|C|$ is the length of the dataset.

Overall, to train our model, we use the following loss function $\mathcal{L}$:

$$\mathcal{L} = \mathcal{L}_{cls} + \gamma \mathcal{L}_f, \tag{15}$$

where γ is a trade-off to balance the importance of the simulation task in the training stage.

3 Experiment setting

Datasets To evaluate the effectiveness of our proposed MMSN for multimodal fact verification with incomplete modality, we choose three public benchmark datasets, FACTIFY (Mishra et al., 2022), MOCHEG (Yao et al., 2023), and Fin-Fact (Rangapur et al., 2023), to conduct experiments. The detailed information on these three datasets can be found in Appendix C.

To imitate real-world scenarios, we mask the textual content of evidence and the visual content of claim and evidence in different proportions, specifically from 0% to 100%[5]. We ensure that at least one of the textual and visual content of evidence is available and manually remove those claim-evidence pairs without both textual and visual content which are regarded as less check-worthy.

[4] Here we only extract entities from claim-evidence pairs with available textual evidence. For those pairs missing evidence text we set the entite set as **0**.

[5] For example, for the 50%-masking setting, we randomly sample 50% claim-evidence pairs and discard some of their modality information adhering to the aforementioned requirements in section 2.1.

Model	FACTIFY						MOCHEG						Fin-Fact					
	0%		**50%**		**100%**		**0%**		**50%**		**100%**		**0%**		**50%**		**100%**	
	Acc	$F1_w$	Acc	$F1_w$	Acc	$F1_w$	Acc	$F1_w$	Acc	$F1_w$	Acc	$F1_w$	Acc	$F1_w$	Acc	$F1_w$	Acc	$F1_w$
DeBERTa (He et al., 2021)	63.61	63.60	58.79	58.77	56.48	56.60	40.74	40.55	35.69	35.86	32.85	32.88	67.32	67.30	64.01	64.00	61.34	61.33
CLIP (Radford et al., 2021)	70.36	70.36	66.23	66.23	63.73	63.71	49.43	49.20	47.88	47.89	44.96	44.89	68.27	68.22	65.73	65.74	62.82	62.82
ConcatNet (Mishra et al., 2022)	73.71	73.64	68.95	68.95	65.45	65.50	50.48	50.12	48.57	48.59	45.56	45.56	68.97	69.02	66.43	66.45	63.03	63.05
PreCoFact (Wang and Peng, 2022)	75.61	75.74	69.53	69.56	66.92	66.91	68.47	68.45	64.38	64.40	60.66	60.66	75.22	75.20	72.79	72.77	69.50	69.50
Logically (Gao et al., 2022)	77.03	77.00	71.35	71.35	67.33	67.30	67.78	67.78	64.98	64.99	61.76	61.74	74.85	74.81	73.14	73.16	70.64	70.68
ECENet (Zhang et al., 2023)	**81.48**	**81.50**	71.45	71.46	68.83	68.82	69.40	69.42	66.42	66.49	63.59	63.61	76.15	76.15	73.41	73.43	71.77	71.77
Multi-KE GAT (Cao et al., 2024)	79.64	79.64	71.38	71.38	68.97	68.97	**70.10**	**70.14**	66.53	66.49	63.48	63.47	76.33	76.33	73.66	73.62	71.55	71.57
DD-IMvMLC-net (Wen et al., 2023)	79.24	79.25	71.65	71.63	69.90	69.94	68.98	68.96	66.54	66.51	64.72	64.71	74.76	74.75	72.36	72.37	70.46	70.46
KDCN (Sun et al., 2023)	80.08	80.08	71.57	71.58	68.87	68.89	68.80	68.83	66.36	66.34	64.46	64.44	76.24	76.27	73.75	73.78	72.06	72.05
MMSN (Ours)	80.73	80.77	**73.60**	**73.61**	**70.98**	**70.99**	70.05	70.09	**66.93**	**66.94**	**65.81**	**65.94**	**77.01**	**77.00**	**74.41**	**74.44**	**73.51**	**73.50**

Table 1: Result of fact verification task with missing modality with different proportions of missing modality. We use weighted F1 ($F1_w$, %) and Accuracy (Acc, %) to evaluate the performance. **Bold** denotes the best performance and underline denotes the second best performance. 50% and 100% denote the proportion of missing modality. In Appendix A, we report the full result in Table 2, 3, and 4.

Baselines To assess the performance of our proposed model, we compare it to several multimodal fact verification approaches. **DeBERTa** (He et al., 2021) leverage the pre-trained model DeBERTa extract textual features to make predictions. **CLIP** (Radford et al., 2021) learns multimodal representations and concatenates them to predict the label. **ConcatNet** (Mishra et al., 2022) utilizes cosine similarity to fuse inner-modal features and concatenates textual and visual representations to obtain multimodal features. **PreCoFact** (Wang and Peng, 2022) uses the co-attention layers to fuse multimodal contents and predict the label. **Logically** (Gao et al., 2022) uses a decision tree classifier with several multimodal features to make predictions. **ECENet** (Zhang et al., 2023) introduces textual and visual entities as external knowledge helping to predict the label. **Multi-KE GAT** (Cao et al., 2024) leverages multi-source knowledge and heterogeneous fusion methods to perform multimodal interactions and make predictions.

Besides, we compare several multimodal representation learning methods with incomplete modality to investigate the effectiveness of our proposed model. **DD-IMvMLC-net** (Wen et al., 2023) uses an encoder-decoder framework to learn comprehensive multimodal representations with missing modality. **KDCN** (Sun et al., 2023) capture the inconsistent information at the cross-modal level and the content-knowledge level to learn a comprehensive multimodal representation with incomplete modality. More detailed descriptions on related work can be found in Appendix D.

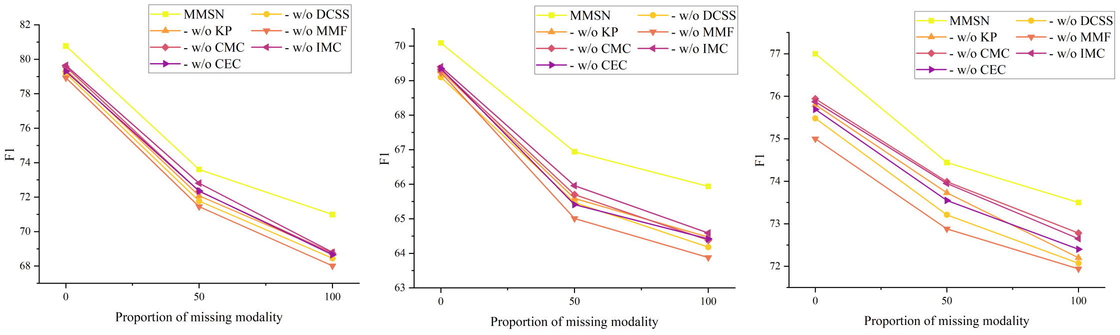
Implementation details We use a Tesla V100-PCIE GPU with 32GB memory for all experiments and implement our model via the Pytorch framework. The seed is set to 43. The number of attention heads is set to 4. The batch size is 32. We set the learning rate as 2e-5. For each claim or evidence, we extract at most 5 key phrases. We employ DeBERTa (He et al., 2021) and Swin Transformer (Liu et al., 2021) as the pre-trained language and visual models. We employ GPT-3.5 as the key-phrase extractor. For γ, we adaptively set to 0.6-0.8 adhering to emperical results. More detailed information is demonstrated in Appendix B.

Evaluation metrics To evaluate the performance of the proposed model, the Accuracy score and the weighted F1 score are used as the evaluation metrics for all three datasets.

4 Result and discussion

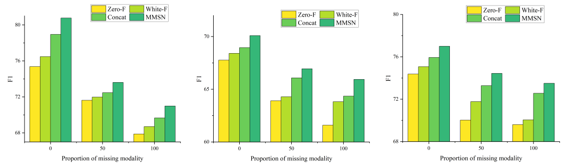
4.1 Overall performance

We conduct the experiments on three datasets and the experimental results are shown in Table 1. Our MMSN achieves comparable performance on all datasets for verification with complete modality information. Here we only report 0%, 50%, and 100% settings, and more specific results are demonstrated in Appendix A. It can be observed that MMSN outperforms other methods in most settings. Traditional multimodal fact verification methods (e.g., ECENet, and MultiKE-GAT) lack flexibility and adaptability to handle the modality-missing scenarios completeness hypothesis. Approaches focusing on multimodal representation learning with missing modalities (e.g., DD-IMvMLC-net and KDCN) improve the capability of generalization but neglect task-specific feature extraction, which leads to low performance. Furthermore, focusing on the 0% setting, the performance of MMSN is comparable to traditional methods. It demonstrates that MMSN can solve data with complete information as well, which further proves the robustness of our model.



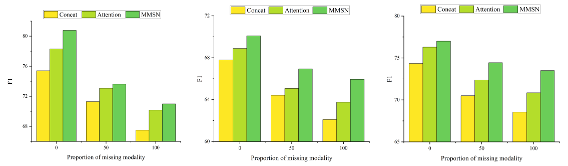
(a) FACTIFY (b) MOCHEG (c) Fin-Fact

Figure 3: The result of ablation studies on 3 datasets. We use weighted F1 ($F1_w$, %) to evaluate the performance. 50% and 100% denote the proportion of missing modality. DCSS, KP, MMF, CMC, IMC, and CEC denote dual-channel soft simulation, fine-grained key phrases, multi-granularity multimodal fusion, cross-modal correlation, intra-modal correlation, and claim-evidence correlation respectively.



(a) FACTIFY (b) MOCHEG (c) Fin-Fact

Figure 4: Experimental results of the analysis of simulation methods on three datasets. We use a weighted F1 score to evaluate the performance.



(a) FACTIFY (b) MOCHEG (c) Fin-Fact

Figure 5: Experimental results of the analysis of fusion methods on three datasets. We use a weighted F1 score to evaluate the performance.

Overall, the experimental results demonstrate that MMSN has the outstanding capability of handling multimodal fact verification tasks. It is more flexible and robust to the real-world circumstances in which news content is missing.

4.2 Ablation study

We conduct the ablation study to analyze key components of MMSN. We remove each component including the dual-channel soft simulation (DCSS), fine-grained key phrases (KP), and multi-granularity multimodal fusion(MMF), respectively. For the DCSS module, we further investigate the importance of cross-modal (CMC), intra-modal (IMC), and claim-evidence (CEC) correlation. The ablation results are shown in Figure 3.

Specifically, intra-modal features are more significant compared to cross-modal simulation, for it contains more crucial features to simulate missing representations. It demonstrates that using only one modality to simulate features is impractical. It can be observed that introducing CMC, IMC, and CEC can also improve the 0% setting performance, which indicates that these fine-grained information and correlations are also crucial to the multimodal fact verification with complete data. Besides, compared to other correlations, claim-evidence correlations have a more crucial impact on the model, which further proves the hypothesis we mention in section 1. Furthermore, the performance degrades the performance significantly on these three datasets in both the 50%-missing and 100%-missing setting, which further indicates that external fine-grained knowledge is beneficial to the label prediction and LLMs are capable of understanding the extraction task and discovering key phrases useful to fact verification. We remove the multi-granularity multimodal fusion module as well. The performance drops rapidly and dramatically, which elucidates that fusion modules are critical to obtaining comprehensive multimodal representations for prediction-making procedures.

Overall, these results of different proportion settings demonstrate the robustness and effectiveness of each component in MMSN.

4.3 Module analysis

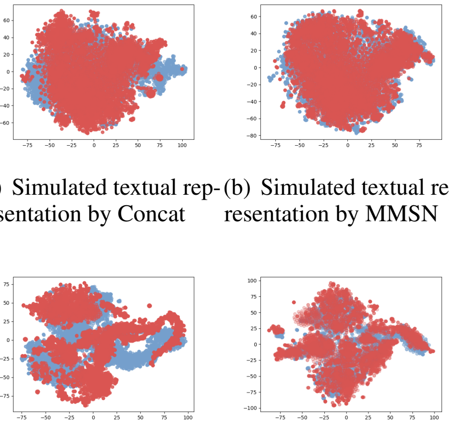
We further conduct several experiments to investigate the usefulness of each proposed module, compared to other available approaches.

Modality simulation We analyze the impact of different simulation methods. We replace the DCSS module with different simulation methods and intend to probe its effectiveness and how better its simulated representations are. We use 3 methods to replace DCSS: (1) zero-filling, which utilizes a zero tensor as a substitution; (2) white-filling, which regards the missing image as a pure white picture; and (3) concatenation, which concatenates the cross-modal and claim-evidence simulated representation rather than in a weighted way. Figure 4 demonstrates the experimental results leveraging these simulation methods. Our model outperforms these simulation methods in weighted F1 scores. It indicates that these approaches omit the significant semantic information contained in visual modalities, which degrades the performance in multimodal tasks. Besides, compared to DCSS, the concatenation method performs worse, which elucidates that simply concatenation cannot capture the comprehensive information between modalities.

Multi-granularity fusion Then, we explore the capability and effectiveness of our proposed multi-granularity multimodal fusion module. We compare our module with two representative-used fusion methods: (1) concatenation-based fusion, simply concatenating all of the multimodal representations; and (2) attention-based fusion, applying the cross-attention mechanism without considering different granularities. Figure 5 shows the experimental results using these three fusion methods. It can be observed that our multi-granularity fusion module outperforms other approaches. It indicates that the main content and the fine-grained knowledge should be treated and operated distinctively to capture both coarse-grained and fine-grained semantic information that is essential to predict the verdict of a claim.

4.4 Simulation analysis

We further compare the quality of simulated features obtained by the dual-channel soft simulation module with the concatenation method and show the result in Figure 6. We can obtain a similar distribution compared to original representations of both textual and visual features, and it illustrates the importance and advantages that cross-modal and claim-evidence correlations are taken into account when we deal with tasks with incomplete modalities.



(a) Simulated textual representation by Concat

(b) Simulated textual representation by MMSN

(c) Simulated visual representation by Concat

(d) Simulated visual representation by MMSN

Figure 6: The comparison of the simulated features and original representations of different methods. The blue dots demonstrate the original representations of evidence images, and the red dots demonstrate the simulated features.

5 Conclusion

This work has investigated the real-world scenario in which multimodal data for fact verification may be incomplete during the limited collection procedure, missing some important multimodal content. We propose a novel Missing Modality-Simulated Network (MMSN) for robust verification. Besides, we design a novel soft simulation module to imitate the missing visual features and eliminate the noise and duplicates by utilizing both intra-modal and inter-modal correlations instead of using only one modality. Moreover, we propose a multi-granular multimodal fusion module to integrate coarse-grained and fine-grained data respectively. The experimental results on three commonly used datasets show that MMSN has been proven to be capable of effectively dealing with modality-missing multimodal fact verification tasks in comparison with other competitive methods.

Limitation

In this paper, there are some limitations that can be improved in future research. First, we do not take LLM-as-verifiers into account. LLMs demonstrate significant performance in NLP tasks, while we only leverage a small part of their capabilities to extract key phrases. How to directly utilize the verification capability of LLMs is still uncovered. Second, we focus on how to simulate the missing information, while there are some approaches to leverage retrieval models to obtain more evidence. They make a compromise between correlations of claim and evidence and information missing problems. This may be one of the future research directions.

References

Saksham Aggarwal, Pawan Kumar Sahu, Taneesh Gupta, and et al. 2022. Gpts at factify 2022: Prompt aided fact-verification (short paper). In *AAAI*.

Mubashara Akhtar, Michael Schlichtkrull, Zhijiang Guo, and et al. 2023. Multimodal automated fact-checking: A survey. In *EMNLP*, pages 5430–5448.

Han Cao, Lingwei Wei, Wei Zhou, and et al. 2024. Multi-source knowledge enhanced graph attention networks for multimodal fact verification. In *ICME*.

Yixuan Chen, Dongsheng Li, Peng Zhang, and et al. 2022. Cross-modal ambiguity learning for multimodal fake news detection. In *WWW*, page 2897–2905.

Abhishek Dhankar, Osmar Zaïane, and François Bolduc. 2022. Uofa-truth at factify 2022 : A simple approach to multi-modal fact-checking. In *AAAI*.

Mudit Dhawan, Shakshi Sharma, Aditya Kadam, and et al. 2023. Game-on: graph attention network based multimodal fusion for fake news detection. *Soc. Netw. Anal. Min.*, page 114.

Jie Gao, Hella-Franziska Hoffmann, Stylianos Oikonomou, and et al. 2022. Logically at factify 2022: Multimodal fact verfication. In *AAAI*.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and et al. 2021. Deberta: Decoding-enhanced bert with disentangled attention. In *ICLR*.

Ruohong Huan, Guowei Zhong, Peng Chen, and et al. 2024. Unimf: A unified multimodal framework for multimodal sentiment analysis in missing modalities and unaligned multimodal sequences. *IEEE TMM*, pages 5753–5768.

Rongrong Ji, Fuhai Chen, Liujuan Cao, and et al. 2019. Cross-modality microblog sentiment prediction via bi-layer multimodal hypergraph learning. *IEEE TMM*, pages 1062–1075.

Mohammadamin Kanaani. 2024. Triple-r: Automatic reasoning for fact verification using language models. In *LREC/COLING*, pages 16831–16840.

Jiho Kim, Sungjin Park, Yeonsu Kwon, and et al. 2023. Factkg: Fact verification via reasoning on knowledge graphs. In *ACL*, pages 16190–16206.

Zheng Lian, Lan Chen, Licai Sun, and et al. 2023. Gcnet: graph completion network for incomplete multimodal learning in conversation. *IEEE TPAMI*.

Weide Liu, Huijing Zhan, Hao Chen, and et al. 2024. Multimodal sentiment analysis with missing modality: A knowledge-transfer approach. *arXiv*.

Ze Liu, Yutong Lin, Yue Cao, and et al. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In *ICCV*.

Shreyash Mishra, Suryavardan S, Amrit Bhaskar, and et al. 2022. FACTIFY: A multi-modal fact verification dataset. In *AAAI*.

Taichi Murayama. 2021. Dataset of fake news detection and fact verification: a survey. *arXiv*.

Peng Qi, Yuyang Zhao, Yufeng Shen, and et al. 2023. Two heads are better than one: Improving fake news video detection by correlating with neighbors. In *Findings of ACL*.

Shuwei Qian and Chongjun Wang. 2023. Com: Contrastive masked-attention model for incomplete multimodal learning. *Neural Networks*, pages 443–455.

Alec Radford, Jong Wook Kim, Chris Hallacy, and et al. 2021. Learning transferable visual models from natural language supervision. In *ICML*, pages 8748–8763.

Aman Rangapur, Haoran Wang, and Kai Shu. 2023. Fin-fact: A benchmark dataset for multimodal financial fact checking and explanation generation. *arXiv*.

Rakibul Hasan Sahar Abdelnabi and Mario Fritz. 2022. Open-domain, content-based, multi-modal fact-checking of out-of-context images via online resources. In *CVPR*, pages 14940–14949.

Jiasheng Si, Yingjie Zhu, and Deyu Zhou. 2023. Exploring faithful rationale for multi-hop fact verification via salience-aware graph learning. In *AAAI*.

Mengzhu Sun, Xi Zhang, Jianqiang Ma, and et al. 2023. Inconsistent matters: A knowledge-guided dual-consistency network for multi-modal rumor detection. *IEEE TKDE*.

Shirin Dabbaghi Varnosfaderani, Canasai Kruengkrai, Ramin Yahyapour, and et al. 2024. Bridging textual and tabular worlds for fact verification: A lightweight, attention-based model. In *LREC/COLING*, pages 2515–2519.

Nguyen Vo and Kyumin Lee. 2020. Where are the facts? searching for fact-checked information to alleviate the spread of fake news. In *EMNLP*, pages 7717–7731.

Qifan Wang, Yinwei Wei, Jianhua Yin, and et al. 2023. Dualgnn: Dual graph neural network for multimedia recommendation. *IEEE TMM*, pages 1074–1084.

Wei-Yao Wang and Wen-Chih Peng. 2022. Team yao at factify 2022: Utilizing pre-trained models and co-attention networks for multi-modal fact verification. In *AAAI*.

Shicai Wei, Chunbo Luo, and Yang Luo. 2023. Mmanet: Margin-aware distillation and modality-aware regularization for incomplete multimodal learning. In *CVPR*, pages 20039–20049.

Jie Wen, Chengliang Liu, Shijie Deng, and et al. 2023. Deep double incomplete multi-view multi-label learning with incomplete labels and missing views. *IEEE Trans. Neural Networks Learn. Syst.*

Xian Xu, Xiao Xu, Xiang Li, and et al. 2023. Grmi: Graph representation learning of multimodal data with incompleteness. In *DASFAA*, pages 286–296.

Barry Menglong Yao, Aditya Shah, Lichao Sun, and et al. 2023. End-to-end multimodal fact-checking and explanation generation: A challenging dataset and models. In *SIGIR*, pages 2733–2743.

Chuanming Yu, Yinxue Ma, Lu An, and et al. 2022. Bcmf: A bidirectional cross-modal fusion model for fake news detection. *Inf. Process. Manag.*, page 103063.

Ziqi Yuan, Yihe Liu, Hua Xu, and et al. 2024. Noise imitation based adversarial training for robust multimodal sentiment analysis. *IEEE TMM*, pages 529–539.

Jiandian Zeng, Jiantao Zhou, and Tianyi Liu. 2023. Robust multimodal sentiment analysis via tag encoding of uncertain missing modalities. *IEEE TMM*, pages 6301–6314.

Congzhi Zhang, Linhai Zhang, and Deyu Zhou. 2024. Causal walk: Debiasing multi-hop fact verification with front-door adjustment. In *AAAI*, pages 19533–19541.

Fanrui Zhang, Jiawei Liu, Qiang Zhang, and et al. 2023. Ecenet: Explainable and context-enhanced network for muti-modal fact verification. In *MM*, pages 1231–1240.

Wanqing Zhao, Yuta Nakashima, Haiyuan Chen, and et al. 2023. Enhancing fake news detection in social media via label propagation on cross-modal tweet graph. In *MM*.

A Overall results

To further analyse the effectiveness of our proposed model, we also conduct extensive experiments to substantiate our conclusion on FACTIFY, MOCHEG, and Fin-Fact. We examine the different proportions of missing information from 0% to 100%. The results are demonstrated in Table 2, 3, and 4.

B Parameter analysis

To further study the impact of the hyper-parameter γ, we conduct several experiments with different settings of γ and report the results in Figure 7.

From Figure 7, at the onset, the model's performance exhibited a positive correlation with γ; however, as the magnitude of a continued to escalate, the model's efficacy showed a subsequent decline. For example, the best performance on FACTIFY is gained when γ is set to 0.6 for 20% percentage of missing modality, and 0.5 for 50%, suggesting the importance of striking a balance between the simulation and prediction objectives to optimize the model's effectiveness. Besides, the different performances between different settings further indicate that MMSN is sensitive to the hyperparameter γ and it is crucial to choose a proper setting for different datasets.

C Dataset statistics

FACTIFY collects multimodal claim-evidence pairs from handles of Indian and US news sources. Each pair in this dataset is classified into five categories, *Support Multimodal, Support Text, Insufficient Multimodal, Insufficient Text*, and *Refute*. **MOCHEG** contains claims associated with politics. Each claim is annotated into three categories, *Supported, Refuted*, and *NEI*. **Fin-Fact** contains claims relevant to financial issues and each claim is categorized into three labels, *True, False*, and *NEI*. The number of these datasets are shown in Table 5.

D Related work

Fact verification aims to predict the verdicts of check-worthy claims with several retrieved evidence. Traditional fact verification approaches only utilize textual information to make predictions (Zhang et al., 2024; Kim et al., 2023; He et al., 2021), which fails to deal with claims with multimodal content. Hence, multimodal fact verification has become a research hotspot. This paper mainly focuses on multimodal fact verification tasks, under the real-world challenge of missing modality. In this section, we will report on the related work in these two research fields.

D.1 Multimodal fact verification

Vo and Lee (2020) draw significant attention to using multimodal content for fact verification by considering both textual and visual content. In recent years, there has been a significant increase in research focusing on multimodal fact verification tasks (Sahar Abdelnabi and Fritz, 2022; Mishra et al., 2022; Wang and Peng, 2022; Zhang et al.,

Model	FACTIFY																					
	0%		10%		20%		30%		40%		50%		60%		70%		80%		90%		100%	
	Acc	$F1_w$	Acc	$F1_w$	Acc	$F1_w$	Acc	$F1_w$	Acc	$F1_w$	Acc	$F1_w$	Acc	$F1_w$	Acc	$F1_w$	Acc	$F1_w$	Acc	$F1_w$	Acc	$F1_w$
DeBERTa (He et al., 2021)	63.61	63.60	62.60	62.56	61.74	61.70	60.88	60.85	59.67	59.67	58.79	58.77	58.31	58.33	58.02	58.00	57.59	57.59	57.05	57.05	56.48	56.60
CLIP (Radford et al., 2021)	70.36	70.36	69.84	69.86	68.92	68.90	68.07	68.07	67.38	67.39	66.23	66.23	65.87	65.60	65.23	65.40	64.95	64.92	64.14	64.33	63.73	63.71
ConcatNet (Mishra et al., 2022)	73.71	73.64	72.96	72.84	71.78	71.50	70.58	70.55	69.66	69.66	68.95	68.95	68.29	68.18	67.73	67.79	66.88	66.80	66.02	66.13	65.45	65.50
PreCoFact (Wang and Peng, 2022)	75.61	75.74	74.64	74.66	73.58	73.40	72.19	72.13	71.00	70.85	69.53	69.56	68.89	68.89	68.32	68.37	67.70	67.75	67.24	67.36	66.92	66.91
Logically (Gao et al., 2022)	77.03	77.00	75.85	75.87	74.69	74.60	73.63	73.44	72.28	72.29	71.35	71.35	70.65	70.61	69.83	69.83	68.94	68.77	68.34	68.33	67.33	67.30
ECENet (Zhang et al., 2023)	**81.48**	**81.50**	78.48	78.46	76.30	76.30	74.12	74.19	73.47	73.49	71.45	71.46	70.35	70.33	69.91	69.93	69.50	69.50	69.04	69.00	68.83	68.82
Multi-KE GAT (Cao et al., 2024)	79.64	79.64	77.88	77.88	75.79	75.75	73.60	73.55	72.18	72.14	71.38	71.38	70.86	70.85	70.11	70.11	69.60	69.57	69.14	69.14	68.97	68.97
DD-IMvMLC-net (Wen et al., 2023)	79.24	79.25	78.40	78.33	76.96	76.93	74.40	74.41	72.45	72.44	71.65	71.63	71.28	71.25	70.86	70.81	70.51	70.50	70.00	70.04	69.90	69.94
KDCN (Sun et al., 2023)	80.08	80.08	**78.77**	**78.78**	76.67	76.62	74.26	74.26	72.58	72.58	71.57	71.58	70.94	70.99	70.20	70.20	69.75	69.77	69.17	69.33	68.87	68.89
MMSN (Ours)	80.73	80.77	78.62	78.61	**77.28**	**77.27**	**75.65**	**75.65**	**74.86**	**74.86**	**73.60**	**73.61**	**72.91**	**72.90**	**72.31**	**72.30**	**71.95**	**71.90**	71.38	71.36	**70.98**	**70.99**

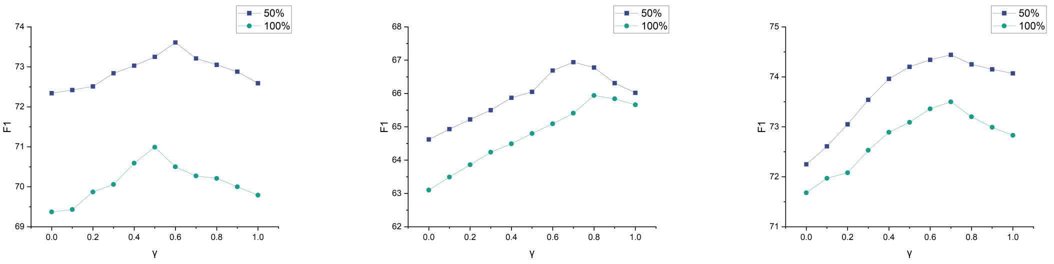
Table 2: Result of fact verification task with missing modality with different proportions of missing modality on FACTIFY dataset. We use weighted F1 ($F1_w$, %) and Accuracy (Acc, %) to evaluate the performance. **Bold** denotes the best performance and underline denotes the second best performance.

Model	MOCHEG																					
	0%		10%		20%		30%		40%		50%		60%		70%		80%		90%		100%	
	Acc	$F1_w$	Acc	$F1_w$	Acc	$F1_w$	Acc	$F1_w$	Acc	$F1_w$	Acc	$F1_w$	Acc	$F1_w$	Acc	$F1_w$	Acc	$F1_w$	Acc	$F1_w$	Acc	$F1_w$
DeBERTa (He et al., 2021)	40.74	40.55	39.70	39.70	38.78	38.78	37.39	37.33	36.65	36.65	35.69	35.86	34.73	34.71	34.14	34.14	33.81	33.86	33.24	33.20	32.85	32.88
CLIP (Radford et al., 2021)	49.43	49.20	48.90	48.89	48.52	48.52	48.29	48.27	48.00	48.00	47.88	47.89	47.38	47.29	46.86	46.82	46.44	46.43	45.71	45.70	44.96	44.89
ConcatNet (Mishra et al., 2022)	50.48	50.12	49.92	49.92	49.67	49.74	49.30	49.26	48.80	48.66	48.57	48.59	48.08	48.08	47.50	47.52	46.94	46.94	46.41	46.38	45.56	45.56
PreCoFact (Wang and Peng, 2022)	68.47	68.45	67.83	67.80	67.04	66.93	66.33	66.32	65.61	65.65	64.38	64.40	63.80	63.87	63.25	63.28	62.40	62.40	61.59	61.58	60.66	60.66
Logically (Gao et al., 2022)	67.78	67.78	66.99	66.93	66.30	66.25	65.89	65.88	65.21	65.20	64.98	64.99	64.00	64.02	63.58	63.58	62.95	62.95	62.47	62.40	61.76	61.74
ECENet (Zhang et al., 2023)	69.40	69.42	68.69	68.69	68.14	68.12	67.85	67.81	67.23	67.22	66.42	66.49	65.60	65.50	65.05	65.03	64.67	64.66	64.29	64.29	63.59	63.61
Multi-KE GAT (Cao et al., 2024)	**70.10**	**70.14**	69.63	69.52	68.98	68.90	68.07	68.06	67.21	67.20	66.53	66.49	65.62	65.60	64.99	64.97	64.56	64.56	63.95	63.90	63.48	63.47
DD-IMvMLC-net (Wen et al., 2023)	68.98	68.96	68.42	68.43	68.20	68.20	67.98	67.98	67.30	67.37	66.54	66.51	66.29	66.29	65.89	65.90	65.52	65.52	65.06	65.02	64.72	64.71
KDCN (Sun et al., 2023)	68.80	68.83	68.45	68.45	68.17	68.11	67.90	67.92	67.20	67.20	66.36	66.34	66.23	66.23	65.90	65.87	65.50	65.41	64.87	64.89	64.46	64.44
MMSN (Ours)	70.05	70.09	**69.89**	**69.82**	**69.17**	**69.17**	**68.66**	**68.73**	**67.68**	**67.68**	**66.93**	**66.94**	**66.30**	**66.30**	**66.19**	**66.22**	**66.07**	**66.07**	**65.96**	**65.99**	**65.81**	**65.94**

Table 3: Result of fact verification task with missing modality with different proportions of missing modality on MOCHEG dataset. We use weighted F1 ($F1_w$, %) and Accuracy (Acc, %) to evaluate the performance. **Bold** denotes the best performance and underline denotes the second best performance.

Model	MOCHEG																					
	0%		10%		20%		30%		40%		50%		60%		70%		80%		90%		100%	
	Acc	$F1_w$	Acc	$F1_w$	Acc	$F1_w$	Acc	$F1_w$	Acc	$F1_w$	Acc	$F1_w$	Acc	$F1_w$	Acc	$F1_w$	Acc	$F1_w$	Acc	$F1_w$	Acc	$F1_w$
DeBERTa (He et al., 2021)	67.32	67.30	66.66	66.67	66.12	66.10	65.54	65.53	64.68	64.65	64.01	64.00	63.58	63.57	62.95	62.96	62.47	62.60	61.89	61.85	61.34	61.33
CLIP (Radford et al., 2021)	68.27	68.22	67.43	67.44	67.15	67.15	66.78	66.74	66.23	66.15	65.73	65.74	65.26	65.27	64.87	64.87	63.99	64.93	63.11	63.1	62.82	62.82
ConcatNet (Mishra et al., 2022)	68.97	69.02	68.50	68.50	67.94	67.98	67.26	67.35	66.84	66.80	66.43	66.45	65.86	65.75	65.27	65.22	64.39	64.39	63.77	63.71	63.03	63.05
PreCoFact (Wang and Peng, 2022)	75.22	75.20	74.83	74.72	74.28	74.26	73.86	73.88	73.17	73.10	72.79	72.77	72.20	72.21	71.77	71.77	71.09	71.04	70.65	70.47	69.50	69.60
Logically (Gao et al., 2022)	74.85	74.81	74.38	74.30	74.12	74.11	73.89	73.84	73.46	73.42	73.14	73.16	72.88	72.88	72.53	72.48	71.95	71.99	71.36	71.35	70.64	70.68
ECENet (Zhang et al., 2023)	76.15	76.15	75.64	75.68	75.00	74.98	74.73	74.73	74.22	74.28	73.41	73.43	73.18	73.15	72.87	72.88	72.54	72.5	72.06	72.01	71.77	71.77
Multi-KE GAT (Cao et al., 2024)	76.33	76.33	75.98	75.98	75.49	75.42	74.37	74.36	73.98	73.86	73.66	73.62	73.14	73.06	72.77	72.75	72.49	72.48	71.96	71.90	71.55	71.57
DD-IMvMLC-net (Wen et al., 2023)	74.76	74.75	74.50	74.50	74.09	74.03	73.82	73.80	73.27	73.26	72.36	72.37	72.05	72.05	71.89	71.83	71.54	71.55	71.00	70.85	70.46	70.46
KDCN (Sun et al., 2023)	76.24	76.27	75.94	75.93	75.68	75.64	75.17	75.17	74.69	74.62	73.75	73.78	73.18	73.06	72.97	72.95	72.68	72.68	72.33	72.33	72.06	72.05
MMSN (Ours)	**77.01**	**77.00**	**76.59**	**76.58**	**76.23**	**76.22**	**75.67**	**75.67**	**75.33**	**75.26**	**74.41**	**74.44**	**74.18**	**74.11**	**73.99**	**73.96**	**73.85**	**73.84**	**73.74**	**73.66**	**73.51**	**73.50**

Table 4: Result of fact verification task with missing modality with different proportions of missing modality on Fin-Fact dataset. We use weighted F1 ($F1_w$, %) and Accuracy (Acc, %) to evaluate the performance. **Bold** denotes the best performance and underline denotes the second best performance.



(a) FACTIFY

(b) MOCHEG

(c) Fin-Fact

Figure 7: Experimental results of the analysis of parameter γ on MOCHEG and Fin-Fact datasets. We use a weighted F1 score to evaluate the performance.

Dataset	Train	Dev	Test
FACTIFY	28,000	7,000	7,500
MOCHEG	11,669	1,490	2,442
Fin-Fact	2,358	336	675

Table 5: The statistics of three datasets.

2023; Cao et al., 2024; Yu et al., 2022; Chen et al., 2022; Dhankar et al., 2022; Aggarwal et al., 2022). By incorporating visual cues, these approaches aim to capture better the semantic meaning of the claim and its supporting evidence, thereby improving the performance of multimodal fact verification systems.

Sahar Abdelnabi and Fritz (2022) used the co-attention mechanism to capture inner-modal features and the CLIP model to obtain inter-modal features. Mishra et al. (2022) introduced a fine-grained fact categorization and an attention-based encoder to extract multimodal representations. Wang and Peng (2022) utilized a co-attention mechanism to integrate textual and visual features. ECENet (Zhang et al., 2023) introduced textual and visual entities as external knowledge helping to predict the label. Cao et al. (2024) introduced multi-source knowledge and constructed a heterogeneous graph for each claim-evidence pair to perform fine-grained and comprehensive multimodal interactions. Yu et al. (2022) considered text-to-image and image-to-text fusion simultaneously and designed a bidirectional fusion network utilizing two separated gating mechanisms to fuse multimodal features bidirectionally. Chen et al. (2022) observed the cross-modal ambiguity in fake statements to learn the ambiguity and difference between modalities, serving as a gating mechanism to control the multimodal fusion level. Dhankar et al. (2022) used cosine similarity to capture inner-modal relations and concatenated representations of both modalities to obtain multimodal features. Due to the satisfactory performance of pre-trained language models like GPT, Aggarwal et al. (2022) leveraged the GPT model and tried to design suitable prompts and verification methods to deal with multimodal fact verification.

D.2 Multimodal representation learning with missing modality

Recent studies on learning high-quality multimodal representation with missing modality tend to utilize simulation-based methods (Qian and Wang, 2023; Zeng et al., 2023; Huan et al., 2024), or to leverage available features to learn modal-invariant representations without simulation (Lian et al., 2023; Xu et al., 2023; Wei et al., 2023).

Qian and Wang (2023) utilized contrastive learning methods to capture shared cross-modal features to learn a better representation of missing modality. Zeng et al. (2023) utilized extra information such as tags to simulate the missing textual and visual information to solve multimodal sentiment classification. UniMF (Huan et al., 2024) introduced a translation module to leverage available information to simulate the missing part for sentiment prediction. Lian et al. (2023) focused on multimodal dialogue systems and utilized a graph-based method to fuse multimodal features and deal with the problem of missing modality. Xu et al. (2023) introduced a bipartite graph structure to capture modal-invariant features. Moreover, Wei et al. (2023) leveraged a teacher model trained with complete data to guide the student model, making predictions with incomplete data.