

Faithful Inference Chains Extraction for Fact Verification over Multi-view Heterogeneous Graph with Causal Intervention

Daoqi Chen¹, Yaxin Li^{1,2}, Zizhong Zhu¹, Xiaowang Zhang^{1*}, Zhiyong Feng¹,

¹College of Intelligence and Computing, Tianjin University

²The Center of National Railway Intelligent Transportation System Engineering and Technology

{chendaoqi, xiaowangzhang}@tju.edu.cn

Abstract

KG-based fact verification verifies the truthfulness of claims by retrieving evidence graphs from the knowledge graph. The *faithful inference chains*, which are precise relation paths between the mentioned entities and evidence entities, retrieve precise evidence graphs addressing poor performance and weak logic for fact verification. Due to the diversity of relation paths, existing methods rarely extract faithful inference chains. To alleviate these issues, we propose Multi-view Heterogeneous Graph with Causal Intervention (MHGCI): (i) We construct a Multi-view Heterogeneous Graph enhancing relation path extraction from the view of different mentioned entities. (ii) We propose a self-optimizing causal intervention model to generate assistant entities mitigating the out-of-distribution problem caused by counterfactual relations. (iii) We propose a grounding method to extract evidence graphs from the KG by faithful inference chains. Experiments on the public KG-based fact verification dataset FactKG demonstrate that our model provides precise evidence graphs and achieves state-of-the-art performance. Our code is available at <https://github.com/CarlosChen1999/MHGCI>.

1 Introduction

With the growing false information, such as unfounded rumors, fake news (Zhou and Zafarani, 2020; Guo et al., 2022), and false generation in large language model (Guan et al., 2024), people’s daily lives and social stability can be profoundly affected. Fact verification (FV) is a valuable task to automatically retrieve evidence in public data to verify claims as SUPPORTED or REFUTED. Most current studies are based on the text or table evidence (Thorne et al., 2018; Aly et al., 2021a), resulting in unreliable reasoning between the claims and the judgments. Kim et al. (2023b) introduce

*Corresponding Author.

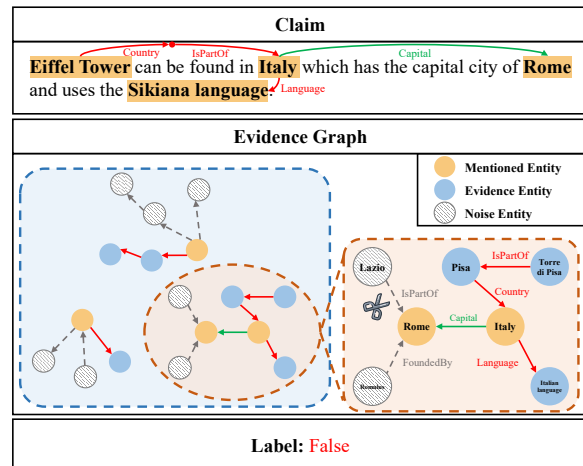


Figure 1: Examples of KG-based fact verification in existing work. **Solid arrows** represent the faithful inference chains. **Green arrows** represent factual relation paths, while **red arrows** represent counterfactual relation paths.

a knowledge graph, structured with factual knowledge as edges and nodes, as evidence for transparent reasoning for the fact verification, shown in Figure 1.

In the current research, the FV task is primarily divided into two stages (Guo et al., 2022): Evidence Retrieval (ER) and Verification Prediction (VP). As a coherent task, the quality of evidence affects the verification prediction and the credibility of the FV system. In previous KG-based evidence retrieval methods, evidence classification models cost substantial computational resources, which classify multi-hop evidence in the KG (Liu et al., 2024). Therefore, retrieval path extraction methods are considered valuable work, which extract relation paths to obtain the evidence graphs. Specifically, these methods extract relation paths like [Language], [Capital] and [City, isPartOf] as retrieval paths for 'Italy' to retrieve evidence graphs in the KG. Therefore, extracting the *faithful inference chains*, which are precise relation paths between

mentioned entities and evidence entities, could obtain precise evidence graphs. Kim et al. (2023b) employs a relation classifier and a hop classifier to extract relation paths on a single entity and the claim. Due to various path directions and lengths, their method fails to extract precise relation paths, leading to noisy evidence graph generation. KG-GPT (Kim et al., 2023a) uses a Large Language Model (LLM) few-shot approach to divide claim sentences by entity pairs and extract relation paths by these segments. Due to training data limitation, the LLM method struggles to extract counterfactual relation paths, which are beyond the real-world distributions. However, the current dataset is hard to contain sufficient counterfactual samples for multi-label classification tasks. Thus, counterfactual relations extraction is considered an out-of-distribution (OOD) problem.

Based on the above findings, we introduce a Multi-view Heterogeneous Graph with causal intervention (MHGCI), a framework to extract *faithful inference chains* in the counterfactual environment. (i) To achieve precise relation path extraction, we construct a Multi-view Heterogeneous Graph (MHG) that incorporates distinct views for different mentioned entities and augments reasoning through assistant entities. (ii) To address the OOD problem caused by counterfactual relations, we propose a self-optimizing causal intervention model for assistant entities. Unlike previous causal intervention methods based on data augmentation (Zhou et al., 2023; Zhu et al., 2023; Lv et al., 2022), our model trains a causal intervention model through invariant risk minimization to remove the spurious association between entities and labels. (iii) Finally, we propose a grounding procedure to transform the relation path into retrieval paths to obtain the evidence graph. Our experiments on the FactKG dataset demonstrate that our model improves the accuracy of fact verification by 4.43% compared to the state-of-the-art methods. For evidence retrieval performance, the precision of relation path extraction reached 74.58%. To summarize, our contributions are as follows:

- We introduce a Multi-view Heterogeneous Graph to ensure precise relation path extraction to reduce noise evidence graphs negative effects.
- We introduce a self-optimizing causal intervention model to address the OOD problem without data augmentation.

- Our experiments demonstrate that our method could provide precise evidence graphs and outperforms the current SOTA methods.

2 Related Work

2.1 Evidence Retrieval for Fact Verification

Evidence retrieval, as a critical part of fact verification, affects the performance and credibility of the fact verification. In this work, the FV task uses various types of knowledge as evidence, such as text, table, and knowledge graph. For text evidence, keyword matching and sentence classification methods are widely used in fact verification (Hanselowski et al., 2018; Zhou et al., 2019; Wan et al., 2021). To obtain precise text evidence, the current research classifies potential evidence texts by the claim, including DQN-base approach (Wan et al., 2021), information bottleneck approach (Paranjape et al., 2020), and graph learning models (Chen et al., 2022). Due to the unique structure, structured evidence poses challenges for retrieval. For table evidence extraction, current research is classified as table-level table extraction (Aly et al., 2021b; Hu et al., 2022) and cell-level evidence extraction (Gi et al., 2021; Acharya, 2021). FactKG (Kim et al., 2023b) is the first dataset for fact verification, utilizing knowledge graphs as evidence. Knowledge graphs involve complex structures, making it challenging to extract effective evidence. Our approach is the first precise evidence extraction in the KG-based fact verification task.

2.2 Causal Inference

The causal inference has been widely used in various fields such as medicine, sociology, and economics for many years (Balke and Pearl, 1995; Richiardi et al., 2013). It provides a method for analyzing data features and estimating potential causal effects to achieve desired objectives. In current work, they apply the Structural Causal Model (SCM) (Schölkopf et al., 2012) in various tasks such as out-of-distribution problem (Lv et al., 2022), graph neural network classification (Wu et al., 2022), and debiased tasks (Zhou et al., 2023; Zhu et al., 2023). The current NLP methods primarily ensure invariant learning across different environments through data augmentation (Zhou et al., 2023; Zhu et al., 2023). However, creating reasonable and sufficient enhancement data for multi-label classification tasks is difficult. In this work, we introduce SCM into fact verification

and propose a self-optimizing causal intervention model based on invariant risk minimization (Arjovsky et al., 2019) to mitigate the OOD problem.

3 Method

In this section, we introduce our framework, Multi-view Heterogeneous Graph with causal intervention (MHGCI) (Figure 2). Firstly, we obtain the entity embedding from claims and construct a multi-view heterogeneous graph (Sec.3.2). Then we use a self-optimizing causal intervention model for assistant entity generation (Sec.3.3) and extract relation paths (Sec.3.4). Finally, we design a grounding method to obtain evidence graphs from the knowledge graph (Sec.3.5).

3.1 Problem Definition

The evidence retrieval task aims to find evidence graphs from the knowledge graph $\mathcal{K} = \{(e_i^s \xrightarrow{r_i} e_i^o) | e_i^s, e_i^o \in \mathcal{E}, r_i \in \mathcal{R}\}_{i=1}^{n_k}$ to verify whether the FV system can SUPPORTED or REFUTED the claim C . In this work, we aim to take the claim C and mentioned entities $\{e_i\}_{i=1}^n$ as input to extract relation paths $\{r_i\}_{i=1}^m$ between the mentioned entities and evidence entities. The mentioned entities e_i and their relation paths $\{r_i\}_{i=1}^m$ are used to construct retrieval paths $\hat{\mathcal{P}}_{e_i} = \{e_i \xrightarrow{r_1} o_1 \dots o_{m-1} \xrightarrow{r_m} o_m | e \in \mathcal{E}, r \in \mathcal{R}\}$ for obtaining evidence graphs $Evi \subset \mathcal{K}$, where o is potential evidence entities.

3.2 Entity Encoding and Graph Construction

Entity Encoding Given a claim $C = \{c_i\}_{i=1}^l$, where l is the length of the claim. First, we add special tokens to represent the sentence feature in C in the format: $C = \{[CLS], c_1, c_2, \dots, c_l\}$. We input C into the pre-trained model BERT (Devlin et al., 2019) to obtain embedding of the claim sentences:

$$H_c = BERT([CLS], c_1, c_2, \dots, c_l) \quad (1)$$

To obtain entity embedding with contextual information, we extract the mentioned entity embedding from the claim embedding. Firstly, The entity position Pos_{e_i} in the claim sentence c is obtained by the longest common substring matching algorithm. Then we obtain the entity embedding based on the position information and smooth it by averaging the entity embedding across different positions. This process is described as follows:

$$Pos_{e_i} = \begin{cases} 1, & \text{where } c[i : i + l_{e_i}] = e_i \\ 0, & \text{otherwise} \end{cases} \quad (2)$$

$$h_{e_i} = \frac{1}{l_{e_i}} \sum_{i=1}^l H_c \circ Pos_{e_i} \quad (3)$$

where l_{e_i} is the length of the entity token in the sentence, \circ denotes element-wise multiplication.

Graph Construction To extract faithful inference chains, we construct a Multi-view Heterogeneous Graph (MHG), as illustrated in Figure 2. According to the different mentioned entities, we construct different views to predict potential relation paths. Specifically, MHG constructs n views, each view consisting of n nodes and $n - 1$ edges. Each view selects a different entity as the central entity. We define each view in MHG: $\mathcal{G}_{e_i} = \{(V_{e_c} \cup \{V_{e_a}\}, E) | e_i = e_c, e_c \neq e_a\}$.

Nodes For each view, it focuses on one mentioned entity, extracting all the relation paths associated with it. According to the role represented in relation extraction, we define two node types: **Central Entity** e_c is a key node in each view. This view is designed to predict all the relation paths associated with it. Each view selects a unique mentioned entity as the central entity. The node embedding is directly derived from the entity embedding. **Assistant Entity** e_a is used to assist relation path extraction around the central entity. For claims with multiple entities, We choose the other mentioned entities, apart from the central entity e_c , as assistant entities e_a . In particular, for claims with a single entity, we use the sentence feature as the assistant entity. Considering the impact of counterfactual relations, we regenerate the assistant entity embedding based on causal intervention. This procedure will be described in the next section.

Edges In this graph, we build the edges as potential relation paths around the central entity. Specifically, the central entity connects to all assistant entities, forming $n - 1$ edges in each view. Each edge represents a relation path extraction problem, which can solve the precise path extraction.

3.3 Assistant Entity Generation

Causal Analyse Counterfactual relations, being beyond real-world distribution, are more likely to become OOD data. Due to the wide range of hypothetical changes in entity pairs, these samples tend to exhibit diverse characteristics to disturb the relation path extraction in OOD data. To address this, we use assistant entities with causal intervention to guide the model to learn similar features. We utilize the Structural Causal Model (SCM)(Schölkopf et al., 2012) to characterize the causal effect of the assistant entities in relation path extraction. As

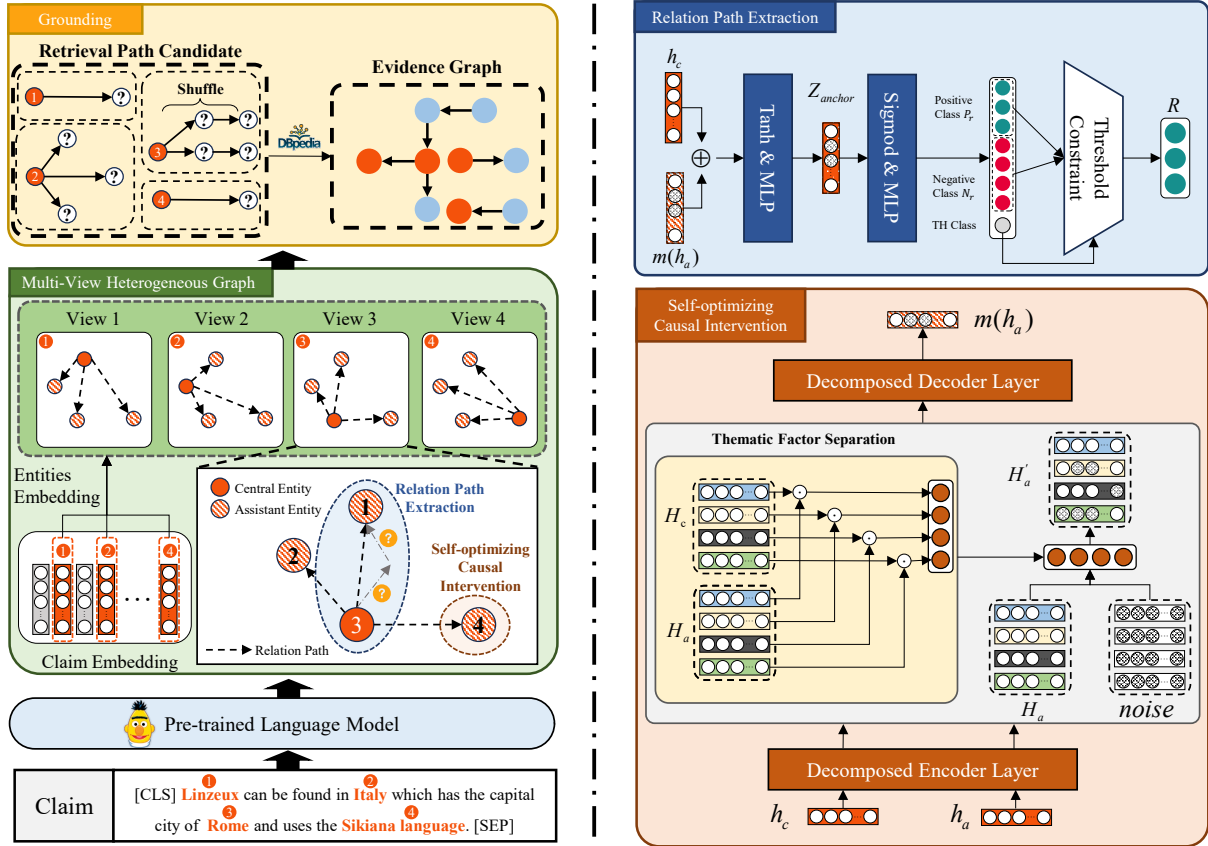


Figure 2: The framework of MHGCI. On the left side, we show the data processing, including constructing the MHG, building the retrieval path, and obtaining the evidence graph. On the right side, we show the main model architecture, including self-optimizing causal intervention and relation path extraction.

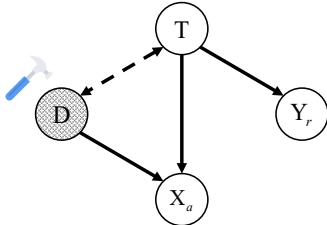


Figure 3: SCM illustrates the causal effect of the assistant entities in relation path extraction. The assistant entity X_a is mixed by thematic factor T and distinctiveness factor D . Only thematic factor T can affect the relation paths extraction.

shown in Figure 3, there are four variables in the SCM: assistant entity features X_a , relation path label Y_r , thematic factor T and distinctiveness factor D . Among these variables, thematic factor T and distinctiveness factor D are latent variables for assistant entity features X_a . For entity pairs within the same relation, entities share some similar features called thematic factor T (e.g., *context information* or *entity type*), while their distinctiveness factor D (e.g., *entity unique feature*) differentiates

them from other entities. In the SCM, the nodes represent these variables, and the edges between nodes denote the causality. Therefore, we will explain the details of the SCM:

- $T \rightarrow X_a \leftarrow D$. The input assistant entity feature X_a is mixed by two factors: thematic factor T and distinctiveness factor D .
- $T \rightarrow Y_r$. In causal analysis, the relation path Y_r is determined by T , which is the common element for the same relation.
- $T \leftrightarrow D$. The double sided arrow indicates the additional probability dependencies (Pearl et al., 2016, 2000) between the thematic factor T and the distinctiveness factor D .

The ideal relation path mining is predicted by T without D . However, $D \leftrightarrow T \rightarrow Y_r$ represents a causal path where D negative impacts on Y_r through T . D creates a spurious association between entities and labels, causing the relation path extraction to fail in the OOD data. We analyze the

cause inference by using Bayes rule:

$$P(Y_r|T) = \sum_D P(Y_r|T, D)P(D|T) \quad (4)$$

To prevent D from creating spurious association, we construct a causal intervention for causal factor D , which will cut the edge $D \leftrightarrow T$ as shown in Figure 3. Based on the backdoor adjustment proposed by Zhu et al. (2023), we use the causal intervention $P(Y_r|do(D))$ to replace the likelihood $P(Y_r|T)$:

$$P(Y_r|do(D = \hat{d})) = \sum_D P(Y_r|D = \hat{d}, T = t)P(T = t) \quad (5)$$

where \hat{d} denotes a causal intervention that ensures $T \rightarrow Y_r$ is invariant across different D . Due to the difficulty of enhancement data for multi-label classification tasks, we propose a self-optimizing causal intervention $m(\cdot)$ to implement $do(\cdot)$. We are inspired by invariant risk minimization (Arjovsky et al., 2019; Chang et al., 2020; Zhou et al., 2023) to implement our causal intervention model. We train $m(\cdot)$ through invariant risk minimization, ensuring that D remains independent of T to avoid spurious association. The goal of the training process is as:

$$\min_{m, f} \mathcal{L}(f(x_c, m(x_a)), y_r), \text{ s.t. } Y_r \perp D | T \quad (6)$$

where $m(\cdot) : D \leftrightarrow T$ mitigates the spurious association resulting from the influence of the distinctiveness factor D on the thematic factor T , $f(\cdot, \cdot) : T \rightarrow Y_r$ denotes the relation paths extraction by T corresponding to Eq.11 and Eq.12, $\mathcal{L}(\cdot, \cdot)$ is the loss function Eq.15.

Self-optimizing Causal Intervention Based on SCM analysis, we propose self-optimizing causal intervention for the assistant entity. To achieve separate the h_T and h_D in h_{e_a} , we map entity embedding into higher-dimensional spaces:

$$H_c = \sigma W_c h_c = [h_c^1, h_c^2, \dots, h_c^k] \quad (7)$$

$$H_a = \sigma W_a h_a = [h_a^1, h_a^2, \dots, h_a^k] \quad (8)$$

where $W_c \in \mathbb{R}^{k \times 1}$, $W_a \in \mathbb{R}^{k \times 1}$ are learnable parameters, σ represents the activation function, k is the dimension of the mapping space. We assume that assistant entity embedding encompasses two dimensions: thematic factor h_T and distinctiveness factor h_D . We posit that h_T should exhibit similarity under the same relation. Therefore, We set this

potential association matrix S_T for the thematic factor T . The S_T calculates attention weight by the entity pairs:

$$S_T = \text{softmax}\left(\frac{H_a H_c^T}{\sqrt{d}}\right) \quad (9)$$

where d is the embedding dimension size. In the S_T , higher scores indicate more thematic factor T that we aim to retain, while lower scores indicate more distinctiveness factor D that we aim to discard. We use random noise following a uniform distribution $U(0, 1)$ to intervene in the distinctiveness factor in higher dimensions, with S_T controlling the extent of noise intervention. Back to Eq.5, we intervene in the embedding of different dimensional features:

$$m(h_a) = \sigma W'_a (S_T \times H_a + (1 - S_T) \times N) \quad (10)$$

where $W'_a \in \mathbb{R}^{1 \times k}$ controls dimension reduction parameters, N is the random noise that intervenes in the distinctiveness factor D .

3.4 Relation Path Extraction

To obtain relation paths, we use the central entity e_c and assistant entity e_a in MHG for extraction. We map the entity pair (e_c, e_a) to the hidden state Z_{anchor} . We assess the correlation likelihood between the relation path r and Z_{anchor} through a learnable linear layer. The process is as follows:

$$Z_{anchor} = \sigma(W_z \text{concat}[h_c, m(h_a)]) \quad (11)$$

$$P(r|e_c, e_a) = \text{sigmoid}(W_r Z_{anchor} + b_r) \quad (12)$$

where $W_z \in \mathbb{R}^{d \times 2d}$, $W_r \in \mathbb{R}^{n_{\mathcal{R}} \times d}$, and $b_r \in \mathbb{R}$ are the model parameters, $n_{\mathcal{R}}$ is the number of relation category. The sigmoid function constrains the output values to $[0, 1]$ as the probability of relation categories.

We use a contrastive learning method (Zhou et al., 2021) to design a path length controller, setting a learnable threshold TH class to filter potential relation paths. Positive classes \mathcal{P}_r are those relation paths with which the central entity is concerned. Negative classes \mathcal{N}_r are those noise relation paths.

$$\mathcal{L}_1 = - \sum_{r \in \mathcal{P}_r} \log \left(\frac{\exp(\text{logit}_r)}{\sum_{r' \in \mathcal{P}_r \cup \{\text{TH}\}} \exp(\text{logit}_{r'})} \right) \quad (13)$$

$$\mathcal{L}_2 = - \log \left(\frac{\exp(\text{logit}_{\text{TH}})}{\sum_{r' \in \mathcal{N}_r \cup \{\text{TH}\}} \exp(\text{logit}_{r'})} \right) \quad (14)$$

$$\mathcal{L} = \mathcal{L}_1 + \mathcal{L}_2 \quad (15)$$

where \mathcal{L}_1 is calculated as the sum of cross-entropy losses for positive classes, which push the positive classes logit higher than the TH class, enabling the extraction of relation paths associated with the central entity. \mathcal{L}_2 computes the cross-entropy loss for negative classes to prevent noise paths in the retrieval path. Two parts are simply summed for the total loss.

3.5 Grounding

Utilizing entity and their relation paths to extract evidence graph from the knowledge graph is defined as 'grounding'. For each view of MHG, the central entity anchor e_c serves as the starting point for evidence retrieval. We obtain a set of relation paths $\{R_1, R_2, \dots, R_{n-1}\}$, where $R_j = \{r_1, r_2, \dots, r_m\}$ is one edge in MHG, n, m is the number and length of relation path. The retrieval path candidates are generated through the random permutations of R_i :

$$p_i = \{e_i, \{shuffle(R_j)\}_{j=1}^{n-1}\} \quad (16)$$

where $shuffle(\cdot)$ represents a method for randomly permuting according to path length. We apply this method for every view of MHG to get the final retrieval path $\hat{\mathcal{P}}$. We use $\hat{\mathcal{P}}$ to obtain evidence graphs $\hat{E}vi$ in \mathcal{K} .

4 Experimental Setup

4.1 Datasets

Type	Written	Colloquial		Total
		Model	Presup	
One-hop	2,106	15,934	1,580	19,530
Conjunction	20,587	15,908	602	37,097
Existence	280	4,060	4,832	9,172
Multi-hop	10,239	16,420	603	27,262
Negation	1,340	12,466	1,807	15,613
Total	34,462	64,788	9,424	108,674

Table 1: Dataset characteristics of FactKG.

FactKG (Kim et al., 2023b) is a large fact verification via knowledge graph dataset. To our knowledge, this dataset is the first to utilize KG evidence to verify unstructured claims. The dataset has 108,674 claims which can be verified by DBpedia (0.1 billion triples) (Lehmann et al., 2015). FactKG possesses a rich variety of claim grammar types, including both written and colloquial language styles. The claims in FactKG are classified into five types, each corresponding to different forms of relation

path reasoning: One-hop, Multi-hop, Conjunction, Existence, and Negation. Specific details are available in the table 1.

4.2 Baseline

In the *claim only* setting, the baseline models only use the claims as input to determine the truthfulness label. We select the three models as baselines: **BlueBERT** (Peng et al., 2019), **BERT** (Devlin et al., 2019) and **Flan-T5** (Chung et al., 2022). While BlueBERT and BERT make predictions based on fully supervised training, Flan-T5 makes predictions using a zero-shot setting. In this setting, we evaluate the performance of pre-trained models that rely solely on claim information and their pre-trained knowledge.

In the *with evidence* setting, the baseline models use both claims and retrieved KG evidence as input to predict the truthfulness label. **KELP** (Liu et al., 2024) constructs a path selection encoder to classify multi-hop evidence for the mentioned entity, obtaining the relevant evidence for LLM-based fact verification. **KG-GPT** (Kim et al., 2023a) is based on a few-shot LLM approach with prompt engineering to divide claim sentences by entity pairs and extract relation paths from these segments, enhancing LLM-based fact verification. **GEAR** (Kim et al., 2023b) is a fully supervised baseline from the FactKG dataset. GEAR employs two independent BERT models to construct retrieval relation paths, using a top-k strategy to retain the most promising candidate paths. Then, they directly employed the fact verification model (Zhou et al., 2019) with the KG setting to verify the truthfulness label.

4.3 Implementation Details

For the evidence retrieval model, MHGCI is implemented based on the Huggingface Transformers framework for the pre-trained model. MHGCI uses the basic model of the hidden dimension 768 dimensions in all experiments. For learning rate scheduling, we respectively set the learning rate for the pre-trained model layers and other layers to $5e-5$ and $1e-4$. MHGCI is optimized using AdamW with a linear warm-up applied for the first 10% of steps. We train the model for 15 epochs with a batch size of 64, selecting the best model after each epoch based on performance on the DEV set. All experiments are conducted on NVIDIA GeForce RTX 4090. The model training cost is approximately 3 hours. For the verification prediction, we remain consistent with the previous

Model	One-hop	Multi-hop	Conjunction	Existence	Negation	Total Accuracy
<i>Fact Verification w/o Evidence</i>						
BlueBERT	60.03	57.79	60.15	59.89	58.90	59.93
BERT	69.64	70.06	63.31	61.84	63.62	65.20
Flan-T5	62.17	69.66	55.29	60.67	55.02	62.70
<i>LLM-Base Fact Verification w/ Evidence</i>						
KG-GPT _{12-shot}	79.10	58.33	76.84	64.00	70.73	71.00
KELP _{12-shot}	-	-	-	-	-	69.20
<i>Fully Supervised Fact Verification w/ Evidence</i>						
GEAR (top-3)	80.88	62.38	83.06	89.20	80.88	78.58
└ (top-5)	80.93	69.00	85.37	83.10	79.57	79.98
└ (top-10)	79.41	72.08	88.43	73.68	74.25	79.68
our model	84.12	72.09	87.88	98.16	85.20	84.41
w/o SCIA	82.50	70.49	86.54	98.05	84.19	83.12
w/o MHG+SCIA	81.24	63.71	70.51	95.98	80.42	75.25

Table 2: Overall performance of FV task and ablation experiment in FactKG dataset. The dataset metrics include sub-task accuracy and total accuracy. The sub-tasks are one-hop, multi-hop, conjunction, existence, and negation.

work GEAR(Kim et al., 2023b). For fairness, we construct the model based on the recommended parameters of the baseline model.

5 Result and Analysis

5.1 Overall Performance

The table 2 shows the overall performance for the fact verification task on the blind TEST set. We evaluate the framework capabilities based on five reasoning sub-tasks. The evaluation metric is the accuracy of the truthfulness label for each claim.

Comparison with the claim only methods. Although BERT exhibits better performance compared to BlueBERT and Flan-T5, indicating that the knowledge embedded in model weights is effective for the reasoning task. However, our model substantially exceeds the claim only baseline by 19.21%. Retrieving evidence is a more effective strategy for the FV task than claim only methods.

Comparison with the LLM-Base method in with evidence setting. For this setting, the FV task performance depends on the evidence retrieval performance and the verification performance. Compared to KG-GPT, a few-shot LLM framework, our model exhibits an 11.73% improvement. Our model consistently exceeds this baseline model across all sub-tasks of reasoning. Due to training cost limitations, we compare the total accuracy with KELP. While KELP improved path selection, their method still performs worse than ours. In summary, the LLM approach does not perform well in KG-based FV tasks.

Model	Setting	Precision	Recall	F1
KG-GPT	top-5	18.400	47.347	26.501
GEAR	top-3	39.502	73.391	51.360
	top-5	25.411	78.685	38.416
	top-10	13.324	82.518	22.944
our model	threshold	74.583	62.728	68.143

Table 3: The evidence retrieval experiment in FactKG. The experiment compared the relation path extraction performance between KG-GPT, GEAR and our model.

Comparison with the fully supervised method in with evidence setting. In this setting, all verification models are consistent, indicating that the performance of the FV task only depends on evidence retrieval performance. For the GEAR model, top-k retrieval strategies produce diverse evidence. As the number of relation paths increases, overall performance generally improves, though this improvement is not unlimited. For the multi-hop and conjunction claims, more paths enhance evidence recall, leading to improved reasoning performance of the model. For one-hop, existence and negation tasks, excessive relation paths decrease model performance, with the maximum impact reaching up to 15.52%. This indicates that extracting faithful inference chains is crucial for optimizing verification performance. Compared to the best results in GEAR, our model shows a 4.43% improvement in accuracy. Our model maintains the best performance in most sub-tasks, especially in tasks with precise path requirements, with the highest metric improvement reaching up to 8.96%. Despite

MHGCI weaker performance in conjunction tasks with a 0.55% decrease, we remain close to the current SOTA level. This demonstrates that the faithful inference chains we extracted effectively support the verification model’s reasoning process.

5.2 Evidence Retrieval Performance

Table 3 presents a comparison experiment of relation path retrieval on the FactKG DEV set¹. We utilize the metrics of Precision, Recall, and F1 score to evaluate the evidence retrieval performance. MHGCI surpasses the GEAR best F1 score by 16.78%. In terms of precision, our model also achieves the best results, surpassing the best performance of the GEAR by 35.08%. It indicates that MHGCI can extract faithful inference chains to obtain precise evidence for the FV task. We observe that as the recall increases, the precision decreases rapidly, which may explain the poor performance of the top-10. We list some retrieval paths and evidence graphs by MHGCI, the case study details are shown in the Sec.5.5.

5.3 Ablation Experiment

To investigate the effectiveness of the MHGCI modules, we construct ablation experiments are shown in Table 2. Firstly, we use the original entity embedding to replace the self-optimizing causal intervention for the assistant entity (SCIA). In the setting without SCIA, the accuracy decreases by 1.29%. The performance decrease across various sub-tasks is similar, indicating that SCIA is effective in all sub-tasks. Then we remove the assistant entity, which causes both MHG and SCIA to malfunction. In the setting without MHG and SCIA, the accuracy drops significantly by 9.16%. This indicates that the assistant entity is crucial. The experiment results show that the modules in MHGCI are effective and necessary.

5.4 Analyzing Counterfactual Influence in Long-Tail Experiment

We observe a training data imbalance across different relation categories in the FactKG dataset. We believe that insufficient training data exacerbates the out-of-distribution (OOD) problem. Therefore, we can design a long-tail experiment to verify our model performance on the OOD problem. We use the FactKG DEV set to design experiments with the fully supervised baseline GEAR_{top-3}. In the Figure

¹The TEST set lacks ground truth for relation path, so we utilize the DEV set to evaluate results.

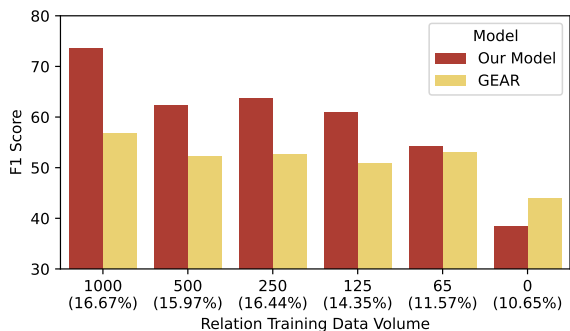


Figure 4: Compare the long-tail relation extraction performance between GEAR_{top-3} and our model. In the figure, the x-axis represents the minimum amount of relation training data volume, with parentheses indicating the proportion of relation categories.

4, we use the F1-score as the metric to observe evidence retrieval performance across different training data volumes. As the training data volume decreases, the performance of the GEAR model continuously decreases, indicating that OOD data can harm their model performance. However, our model exhibits a different performance. Although our model performance also declines, the rate of degradation is slower than that of the GEAR model. When the training data volume exceeds 65, our model maintains the highest performance, indicating that it effectively mitigates the OOD problem in the evidence retrieval task. Especially, when the training data volume is below 65, our model F1-score is worse than GEAR. This may result from the inability to identify entity embedding spaces for causal interventions with low training data volume, which affects the thematic factors.

5.5 Case Study

In Figure 5, we present several retrieval path examples and evidence graph cases obtained by the MHGCI framework. In the one-hop task, MHGCI can retrieve faithful inference chains even when the relation is counterfactual. In the conjunction task, MHGCI can retrieve faithful inference chains across multiple directions. In the multi-hop task, MHGCI can retrieve multi-hop relation paths with precise hops and classifications. Although many noisy retrieval paths are generated due to the shuffle method, this approach is necessary since the relation order cannot be derived directly from the claim. For the existence and negative tasks, the differences lie in the claim style rather than the evidence style, so they are similar to these tasks.


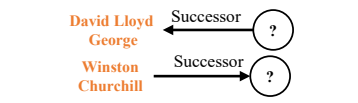
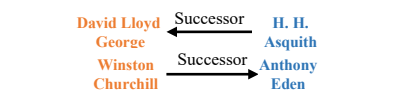

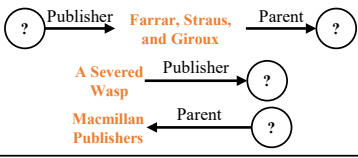
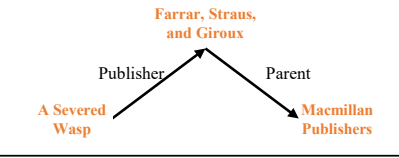

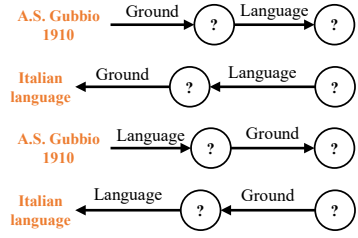
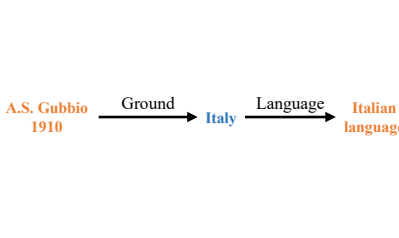
Claim	Retrieval Path	Evidence Graph
 <p>David Lloyd George was Winston Churchill's successor!</p> <p>Reasoning Type: One-hop</p>		
 <p>A Severed Wasp was published by Farrar, Straus, and Giroux, a subsidiary of Macmillan Publishers.</p> <p>Reasoning Type: Conjunction</p>		
 <p>Well A.S. Gubbio 1910 is located in a country where the Italian language is spoken.</p> <p>Reasoning Type: Multi-hop</p>		

Figure 5: Several MHGCI retrieval cases on FactKG. The **orange font** represents the mentioned entities, while the **blue font** represents the evidence entities in the KG.

6 Conclusion

In this work, we propose Multi-view Heterogeneous Graph with Causal Intervention (MHGCI) to extract faithful inference chains for fact verification. Specifically, MHG achieves precise relation path extraction through a multi-view heterogeneous graph. In addition, this method effectively addresses the OOD problem caused by counterfactual relations through self-optimizing causal intervention. In summary, our model provides precise evidence graphs for fact verification. Experiments on the FactKG dataset demonstrate the MHGCI outperforms all baseline models.

Limitations

The MHGCI introduces a method for extracting faithful inference chains, reducing the volume of evidence graphs. In the evidence retrieval, we observed that MHGCI recall is lower than the SOTA method. This could lead to missing crucial evidence, resulting in overall reasoning errors in fact verification. In addition, we observe that the MHGCI becomes ineffective when the training volume is below 65 instances. The added noise disrupts the original effective embeddings, showing a limitation of this approach.

Acknowledgments

This work was supported by Open Project of The Center of National Railway Intelligent Transporta-

tion System Engineering and Technology, CHINA ACADEMY OF RAILWAY SCIENCES CORPORATION LIMITED (RITS2023KF05).

References

- Kaushik Acharya. 2021. [KaushikAcharya at SemEval-2021 task 9: Candidate generation for fact verification over tables](#). In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 1271–1275, Online. Association for Computational Linguistics.
- Rami Aly, Zhijiang Guo, Michael Sejr Schlichtkrull, James Thorne, Andreas Vlachos, Christos Christodoulopoulos, Oana Cocarascu, and Arpit Mittal. 2021a. [The fact extraction and VERification over unstructured and structured information \(FEVEROUS\) shared task](#). In *Proceedings of the Fourth Workshop on Fact Extraction and VERification (FEVER)*, pages 1–13, Dominican Republic. Association for Computational Linguistics.
- Rami Aly, Zhijiang Guo, Michael Sejr Schlichtkrull, James Thorne, Andreas Vlachos, Christos Christodoulopoulos, Oana Cocarascu, and Arpit Mittal. 2021b. [The fact extraction and VERification over unstructured and structured information \(FEVEROUS\) shared task](#). In *Proceedings of the Fourth Workshop on Fact Extraction and VERification (FEVER)*, pages 1–13, Dominican Republic. Association for Computational Linguistics.
- Martín Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. 2019. [Invariant risk minimization](#). *CoRR*, abs/1907.02893.

- Alexander Balke and Judea Pearl. 1995. [Counterfactuals and policy analysis in structural models](#). In *UAI '95: Proceedings of the Eleventh Annual Conference on Uncertainty in Artificial Intelligence, Montreal, Quebec, Canada, August 18-20, 1995*, pages 11–18. Morgan Kaufmann.
- Shiyu Chang, Yang Zhang, Mo Yu, and Tommi Jaakkola. 2020. [Invariant rationalization](#). In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 1448–1458. PMLR.
- Zhendong Chen, Siu Cheung Hui, Fuzhen Zhuang, Lejian Liao, Fei Li, Meihuizi Jia, and Jiaqi Li. 2022. [Evidencenet: Evidence fusion network for fact verification](#). In *Proceedings of the ACM Web Conference 2022*, WWW '22, page 2636–2645, New York, NY, USA. Association for Computing Machinery.
- Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. [Scaling instruction-finetuned language models](#). *Preprint*, arXiv:2210.11416.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. [BERT: Pre-training of deep bidirectional transformers for language understanding](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.
- In-Zu Gi, Ting-Yu Fang, and Richard Tzong-Han Tsai. 2021. [Verdict inference with claim and retrieved elements using RoBERTa](#). In *Proceedings of the Fourth Workshop on Fact Extraction and VERification (FEVER)*, pages 60–65, Dominican Republic. Association for Computational Linguistics.
- Xinyan Guan, Yanjiang Liu, Hongyu Lin, Yaojie Lu, Ben He, Xianpei Han, and Le Sun. 2024. [Mitigating large language model hallucinations via autonomous knowledge graph-based retrofitting](#). *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(16):18126–18134.
- Zhijiang Guo, Michael Schlichtkrull, and Andreas Vlachos. 2022. [A survey on automated fact-checking](#). *Transactions of the Association for Computational Linguistics*, 10:178–206.
- Andreas Hanselowski, Hao Zhang, Zile Li, Daniil Sorokin, Benjamin Schiller, Claudia Schulz, and Iryna Gurevych. 2018. [UKP-athene: Multi-sentence textual entailment for claim verification](#). In *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*, pages 103–108, Brussels, Belgium. Association for Computational Linguistics.
- Nan Hu, Zirui Wu, Yuxuan Lai, Xiao Liu, and Yansong Feng. 2022. [Dual-channel evidence fusion for fact verification over texts and tables](#). In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5232–5242, Seattle, United States. Association for Computational Linguistics.
- Jiho Kim, Yeonsu Kwon, Yohan Jo, and Edward Choi. 2023a. [KG-GPT: A general framework for reasoning on knowledge graphs using large language models](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 9410–9421, Singapore. Association for Computational Linguistics.
- Jiho Kim, Sungjin Park, Yeonsu Kwon, Yohan Jo, James Thorne, and Edward Choi. 2023b. [FactKG: Fact verification via reasoning on knowledge graphs](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16190–16206, Toronto, Canada. Association for Computational Linguistics.
- Jens Lehmann, Robert Isele, Max Jakob, Anja Jentzsch, Dimitris Kontokostas, Pablo N. Mendes, Sebastian Hellmann, Mohamed Morsey, Patrick van Kleef, Sören Auer, and Christian Bizer. 2015. [Dbpedia - A large-scale, multilingual knowledge base extracted from wikipedia](#). *Semantic Web*, 6(2):167–195.
- Haochen Liu, Song Wang, Yaochen Zhu, Yushun Dong, and Jundong Li. 2024. [Knowledge graph-enhanced large language models via path selection](#). In *Findings of the Association for Computational Linguistics ACL 2024*, pages 6311–6321, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.
- Fangrui Lv, Jian Liang, Shuang Li, Bin Zang, Chi Harold Liu, Ziteng Wang, and Di Liu. 2022. [Causality inspired representation learning for domain generalization](#). In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8036–8046.
- Bhargavi Paranjape, Mandar Joshi, John Thickstun, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2020. [An information bottleneck approach for controlling conciseness in rationale extraction](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1938–1952, Online. Association for Computational Linguistics.
- Judea Pearl, Madelyn Glymour, and Nicholas P Jewell. 2016. [Causal inference in statistics: A primer](#). 2016. *Internet resource*.
- Judea Pearl et al. 2000. [Models, reasoning and inference](#). *Cambridge, UK: Cambridge University Press*, 19(2):3.

- Yifan Peng, Shankai Yan, and Zhiyong Lu. 2019. [Transfer learning in biomedical natural language processing: An evaluation of BERT and ELMo on ten benchmarking datasets](#). In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 58–65, Florence, Italy. Association for Computational Linguistics.
- Lorenzo Richiardi, Rino Bellocco, and Daniela Zugna. 2013. [Mediation analysis in epidemiology: methods, interpretation and bias](#). *International Journal of Epidemiology*, 42(5):1511–1519.
- Bernhard Schölkopf, Dominik Janzing, Jonas Peters, Eleni Sgouritsa, Kun Zhang, and Joris M. Mooij. 2012. [On causal and anticausal learning](#). In *Proceedings of the 29th International Conference on Machine Learning, ICML 2012, Edinburgh, Scotland, UK, June 26 - July 1, 2012*. icml.cc / Omnipress.
- James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. [FEVER: a large-scale dataset for fact extraction and VERification](#). 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 809–819, New Orleans, Louisiana. Association for Computational Linguistics.
- Hai Wan, Haicheng Chen, Jianfeng Du, Weilin Luo, and Rongzhen Ye. 2021. [A DQN-based approach to finding precise evidences for fact verification](#). In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1030–1039, Online. Association for Computational Linguistics.
- Yingxin Wu, Xiang Wang, An Zhang, Xiangnan He, and Tat-Seng Chua. 2022. [Discovering invariant rationales for graph neural networks](#). In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.
- Fan Zhou, Yuzhou Mao, Liu Yu, Yi Yang, and Ting Zhong. 2023. [Causal-debias: Unifying debiasing in pretrained language models and fine-tuning via causal invariant learning](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4227–4241, Toronto, Canada. Association for Computational Linguistics.
- Jie Zhou, Xu Han, Cheng Yang, Zhiyuan Liu, Lifeng Wang, Changcheng Li, and Maosong Sun. 2019. [GEAR: Graph-based evidence aggregating and reasoning for fact verification](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 892–901, Florence, Italy. Association for Computational Linguistics.
- Wenxuan Zhou, Kevin Huang, Tengyu Ma, and Jing Huang. 2021. [Document-level relation extraction with adaptive thresholding and localized context pooling](#). *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(16):14612–14620.
- Xinyi Zhou and Reza Zafarani. 2020. [A survey of fake news: Fundamental theories, detection methods, and opportunities](#). *ACM Comput. Surv.*, 53(5).
- Jiazheng Zhu, Shaojuan Wu, Xiaowang Zhang, Yuexian Hou, and Zhiyong Feng. 2023. [Causal intervention for mitigating name bias in machine reading comprehension](#). In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 12837–12852, Toronto, Canada. Association for Computational Linguistics.

A Cross Styles of Claim Experiment

Input Type	Model	$W \rightarrow W$	$W \rightarrow C$	$C \rightarrow C$	$C \rightarrow W$
Claim Only	BlueBERT	64.76	56.28	58.77	63.92
	BERT	71.75	63.85	68.10	69.43
With Evidence	GEAR	81.00	75.43	80.81	78.80
	Our Model	83.76	75.48	84.38	83.73

Table 4: Diversity cross style claim experiments for MHGCI. W refers to the written style claims and C refers to the colloquial style claims. $W \rightarrow C$ means that only train by written styles sub-dataset to predict the colloquial style claims.

Table 4 analyzes the performance of our model across various styles of claims. We divided the dataset into two disjoint parts based on the grammatical style of the claims: written style and colloquial style. We trained the model on different sub-datasets to validate its generalization ability across various types of claims. Our model achieved the best results across all settings, with an average increase of 2.83% compared to the SOTA. In addition, we also discovered some interesting variations in the experimental data. In different styles of claim, the model tends to adapt to the training language style, leading to performance degradation in cross-style settings. However, in the $C \rightarrow W$ setting, we noticed that the *claim only* model would unexpectedly show an improvement in the experiment. Hence, we think colloquial style claims are more prone to causing performance bottlenecks, which can be alleviated by integrating colloquial style corpora during training. This viewpoint is further supported by the $W \rightarrow C$ experiments, where our model performance decrease is more than the baseline, compared to the $W \rightarrow W$.