MEVER: Multi-Modal and Explainable Claim Verification with Graph-based Evidence Retrieval

Delvin Ce Zhang[1], Suhan Cui[2], Zhelin Chu[3], Xianren Zhang[4], Dongwon Lee[5]

[1]University of Sheffield, [2]University of Science and Technology Beijing,
[3]University of California San Diego, [4,5]The Pennsylvania State University

[1]delvin.ce.zhang@sheffield.ac.uk, [2]suhan@ustb.edu.cn,
[3]12111105@mail.sustech.edu.cn, [4,5]{xzz5508,dongwon}@psu.edu

Abstract

Verifying the truthfulness of claims usually requires joint multi-modal reasoning over both textual and visual evidence, such as analyzing both textual caption and chart image for claim verification. In addition, to make the reasoning process transparent, a textual explanation is necessary to justify the verification result. However, most claim verification works mainly focus on the reasoning over textual evidence only or ignore the explainability, resulting in inaccurate and unconvincing verification. To address this problem, we propose a novel model that jointly achieves evidence retrieval, multi-modal claim verification, and explanation generation. For evidence retrieval, we construct a two-layer multi-modal graph for claims and evidence, where we design image-to-text and text-to-image reasoning for multi-modal retrieval. For claim verification, we propose token- and evidence-level fusion to integrate claim and evidence embeddings for multi-modal verification. For explanation generation, we introduce multi-modal Fusion-in-Decoder for explainability. Finally, since almost all the datasets are in general domain, we create a scientific dataset, AIChartClaim, in AI domain to complement claim verification community. Experiments show the strength of our model. Code and datasets are available at https://github.com/cezhang01/mever.

1 Introduction

The dissemination of erroneous claims and findings can mislead researchers and the public, leading to unnecessary concerns. This underscores the urgent need for developing a claim verification method to automatically assess the truthfulness of the claims.

Existing methods primarily rely on textual evidence for claim verification. However, with the advent of multimedia, many claims are derived from various types of data, and simply reasoning over textual evidence is insufficient for accurate verification. For example, Fig. 1 illustrates a scientific

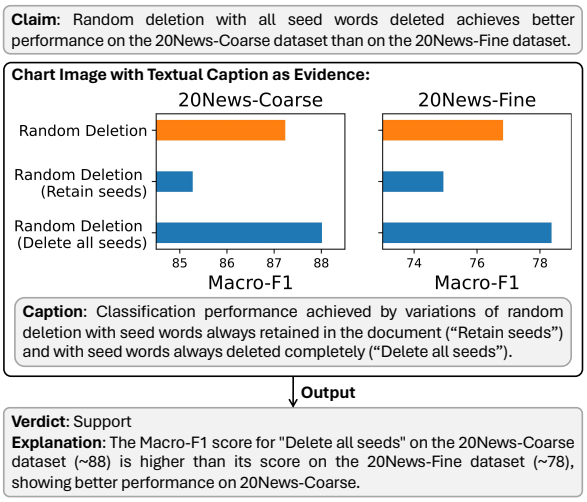


Figure 1: Illustration of multi-modal and explainable claim verification, taken from AIChartClaim dataset.

claim concluded from chart images with textual caption. To verify the claim, we need to unify both textual and visual evidence for reasoning and verification. Meanwhile, to make the reasoning process transparent, an additional textual explanation is necessary to justify the ruling process and verification result. However, most existing works mainly focus on the reasoning over textual evidence only (Liu et al., 2020) or ignore the explainability of the ruling process (Zhang et al., 2024b), leading to inaccurate and unconvincing verification.

Challenges and Approach. To overcome these limitations, we propose MEVER, **M**ulti-Modal and **E**xplainable Claim **V**erification with **E**vidence **R**etrieval, to address below three open questions.

First, *how to integrate both textual and visual evidence for multi-modal evidence retrieval?* Some works, e.g., MultiVerS (Wadden et al., 2022) and DECKER (Zou et al., 2023), rely on external tools (Robertson et al., 2009) for evidence retrieval. Other works, e.g., JustiLM (Zeng and Gao, 2024) and RAV (Zheng et al., 2024), design an in-built text-only retriever. However, they are uni-modal

Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics
Volume 1: Long Papers, pages 5236–5255
March 24-29, 2026 ©2026 Association for Computational Linguistics

and ignore visual evidence, leading to inaccurate retrieval. We design a multi-modal evidence retriever. We construct a two-layer multi-modal graph for claims and evidence. Image-to-text and text-to-image reasoning integrates multi-modal data into claim and evidence embeddings for retrieval.

Second, *how to reason between multi-modal claims and evidence for multi-modal verification?* Some models, e.g., GEAR (Zhou et al., 2019) and Transformer-XH (Zhao et al., 2020), integrate textual claims and evidence for verification, failing to capture visual signals. Recently, multi-modal methods are proposed, e.g., Mocheg (Yao et al., 2023) and ESCNet (Zhang et al., 2024b). However, these methods integrate claims with textual and visual evidence separately, failing to aggregate both textual and visual evidence for joint reasoning. As shown in Fig. 1, images and texts provide complementary information, and jointly aggregating both could reveal insightful discovery. In our model, we design token-level multi-modal fusion and evidence-level hierarchical fusion to achieve fine-grained cross-modal information exchange between claims and evidence. Eventually, we obtain unified claim and evidence embeddings for multi-modal verification.

Third, *how to leverage multi-modal data for explanation generation?* Most explainable models leverage textual claims and evidence to generate explanation, e.g., JustiLM (Zeng and Gao, 2024) and Mocheg (Yao et al., 2023). We design a multi-modal Fusion-in-Decoder module, which generates explanation by capturing both multi-modal and multiple pieces of evidence. Besides, to make the explanation consistent with the verification result, we further design a consistency regularizer.

Besides, since most multi-modal datasets are in general domain, we create a scientific dataset in AI domain to complement research community. Our dataset, AIChartClaim in Fig. 1, contains scientific discoveries as claims, chart images with textual captions as evidence, and explanations. It is necessary to have such scientific dataset, since understanding increases and decreases in quantities in charts with scientific language is a crucial reasoning ability for language models. Though ChartCheck (Akhtar et al., 2024) and ChartFC (Akhtar et al., 2023) are chart datasets, their content is still in general domain and does not discuss scientific concepts. See Table 1 for comparison to selected datasets.

Contributions. *First*, we propose a novel model, MEVER, consisting of multi-modal evidence retrieval, claim verification, and explanation generation. Specifically, we design a two-layer multi-modal graph for evidence retrieval. *Second*, to fully exchange multi-modal information between claims and evidence for verification, we design token- and evidence-level fusion. *Third*, to make the ruling process transparent, we design multi-modal Fusion-in-Decoder and a consistency regularizer for explanation generation. *Fourth*, we create a scientific dataset to complement research community. Experiments verify the strength of our model.

2 Related Work

Claim verification. Previous works are text-only, GEAR (Zhou et al., 2019), KGAT (Liu et al., 2020), TransformerXH (Zhao et al., 2020), MultiVerS (Wadden et al., 2022), HESM (Subramanian and Lee, 2020), DREAM (Zhong et al., 2020), CausalWalk (Zhang et al., 2024a), ProgramFC (Pan et al., 2023), CareerScape (Yamashita et al., 2025), UKE (Wu et al., 2024a), and AKEW (Wu et al., 2024b). They ignore visual evidence, leading to inaccurate result. Multi-modal methods include Mocheg (Yao et al., 2023), ESCNet (Zhang et al., 2024b), CCN (Abdelnabi et al., 2022), MR2Retrived (Hu et al., 2023), CutPaste&Find (Nguyen et al., 2025a), etc. These methods, except Mocheg, focus on verification, and ignore evidence retrieval or explainability. Our model consists of evidence retrieval, claim verification, and explanation generation.

Retrieval-augmented verification conducts evidence retrieval and claim verification jointly with Retrieval-Augmented Generation (Izacard et al., 2023), such as JustiLM (Zeng and Gao, 2024), RAV (Zheng et al., 2024), RAFTS (Yue et al., 2024), ARSJoint (Zhang et al., 2021), etc. They consider textual evidence only. We construct a two-layer graph for multi-modal evidence retrieval and claim verification.

Explainable claim verification produces textual explanation to justify the verification result (Atanasova, 2024; Kotonya and Toni, 2020a; Zhao et al., 2024; Russo et al., 2023; Yao et al., 2023). They are text-only and do not capture images. Our model designs a multi-modal Fusion-in-Decoder to incorporate multi-modal data for explainability.

Scientific and chart claim verification trains models on scientific and chart claims. SciFact (Wadden et al., 2020) is a scientific text-based dataset with no images or explanation. ChartCheck (Akhtar et al., 2024) and ChartFC (Akhtar et al., 2023) are chart datasets in general domain. Ours

Table 1: Comparison between selected datasets.

Dataset	Multi-Modal	Explainable	Scientific	Source
FEVER				Wiki
FEVEROURS				Wiki
PUBHEALTH		✓		FACTWeb
SciFact			✓	Paper
BearFact			✓	Paper
Check-COVID			✓	Paper
FACTIFY	✓			FACTWeb
NewsCLIPpings	✓			Internet
MR2	✓			Internet
MMCV	✓			MultimodalQA
ChartFC	✓			Internet
ChartCheck	✓	✓		Internet
Mocheg	✓	✓		FACTWeb
AIChartClaim	✓	✓	✓	Paper

is a multi-modal dataset with explanation in scientific domain. Table 1 shows comparison to selected data (Thorne et al., 2018; Aly et al., 2021; Kotonya and Toni, 2020b; Wuehrl et al., 2024; Wang et al., 2023; Mishra et al., 2022; Luo et al., 2021; Hu et al., 2023; Wang et al., 2025). Existing survey (Huang et al., 2024) summarizes more chart works, e.g., ChartT5 (Zhou et al., 2023), MatCha (Liu et al., 2023), MMC (Liu et al., 2024a), ChartAssistant (Meng et al., 2024).

Graph learning aims to capture both graph structure and node attributes for graph-based tasks (Hamilton et al., 2017; Zhang and Lauw, 2020, 2023; Zhang et al., 2023a; Nguyen et al., 2025b; Luo et al., 2025). However, most of them are designed for self-supervised learning or for supervised generation. Our work focuses on using graph structure to assist multi-modal claim verification.

3 Model Architecture

We introduce MEVER, Multi-Modal and Explainable Claim Verification with Evidence Retrieval. Table 8 summarizes math notations.

3.1 Problem Formulation

We have a multi-modal dataset $\mathcal{D} = \{\mathcal{C}, \mathcal{T}, \mathcal{I}, \mathcal{E}\}$. $\mathcal{C} = \{c_n\}_{n=1}^{N}$ is a set of N claims. Textual evidence set $\mathcal{T} = \{t_m\}_{m=1}^{T}$ is a corpus of T evidence texts. Each evidence text t is associated with a set of images $\mathcal{I}(t) = \{i_t\} \subset \mathcal{I}$ where $\mathcal{I}$ is an image set. An evidence text may have multiple images, e.g., a caption with multiple charts in Fig. 1. Sometimes, claim c also has images $\mathcal{I}(c) = \{i_c\} \subset \mathcal{I}$. $\mathcal{E} = \{e_n\}_{n=1}^{N}$ is a set of N explanations for N corresponding claims. If we do not observe the association between texts and images, we use pretrained CLIP (Radford et al., 2021) for alignment.

Given $\mathcal{C}$, $\mathcal{T}$, and $\mathcal{I}$ as inputs, our model uses textual evidence $\mathcal{T}$ and visual evidence $\mathcal{I}$ to verify claims $\mathcal{C}$. For each claim c, we have two outputs. One is claim's veracity label $\hat{y} \in \mathcal{Y} = \{\text{SUPPORT, REFUTE, NEI}\}$, i.e., whether the evidence supports, refutes, or does not have enough information to verify the claim. The other is explanation $\hat{e}$ for reasoning process. Note that some datasets (Akhtar et al., 2024) do not have NEI label.

Fig. 2 shows the overall model architecture with three main modules, (a-c) evidence retrieval, (d) claim verification, and (e) explanation generation.

3.2 Evidence Retrieval with Two-Layer Multi-Modal Graph

Graph construction. For each evidence text $t \in \mathcal{T}$ and its associated images $\mathcal{I}(t) \subset \mathcal{I}$, we construct a two-layer multi-modal graph in Fig. 2(a-b). The bottom layer is textual layer, and each node is an evidence text. The top layer is visual layer, and each node is an image. Cross-layer edges denote the correspondence between text t and its images $\mathcal{I}(t)$. We also add intra-layer edges. For visual layer, we add fully connected edges among images of the same text for multi-image reasoning. For textual layer, we add self-loop edge. During evidence retrieval, we treat each evidence independently, thus we add self-loop edge. During verification in Sec. 5.2, we will add fully connected edges among multiple retrieved evidence for multi-evidence reasoning.

Image-to-text reasoning. Evidence reasoning consists of image-to-text and text-to-image reasoning. We first present image-to-text reasoning here.

For each evidence text t, we use $\mathbf{H}_t^{(l)} = [\mathbf{h}_{t,\text{CLS}}^{(l)}, \mathbf{h}_{t,1}^{(l)}, \mathbf{h}_{t,2}^{(l)}, ...]$ to represent the output from the l-th Transformer step. Here we use "step" to replace "layer" (Vaswani et al., 2017) to differentiate from our two-layer graph. Similarly, for each evidence image i, we have output from the l-th ViT step $\mathbf{Z}_i^{(l)} = [\mathbf{z}_{i,\text{CLS}}^{(l)}, \mathbf{z}_{i,1}^{(l)}, \mathbf{z}_{i,2}^{(l)}, ...]$. Since an evidence text may have multiple images, we use Graph Neural Network (GNN) (Hamilton et al., 2017) to aggregate multiple images. We first project text and image embeddings into the same space.

$$\tilde{\mathbf{h}}_{t,\text{CLS}}^{(l)} = \mathbf{W}_{\text{txt}}\mathbf{h}_{t,\text{CLS}}^{(l)}, \quad \tilde{\mathbf{z}}_{i,\text{CLS}}^{(l)} = \mathbf{W}_{\text{img}}\mathbf{z}_{i,\text{CLS}}^{(l)}. \quad (1)$$

$\mathbf{W}_{\text{txt}}, \mathbf{W}_{\text{img}} \in \mathbb{R}^{d \times d}$ are projection matrices for text and image, respectively. We use [CLS] tokens as text and image embeddings. We design image-to-text attention, shown by green arrows in Fig. 2(a).

$$a_{t,i} = \text{softmax}\Big(\text{sigmoid}(\mathbf{b}_{\text{i2t}}^{\top}[\tilde{\mathbf{h}}_{t,\text{CLS}}^{(l)} || \tilde{\mathbf{z}}_{i,\text{CLS}}^{(l)}])\Big) \quad (2)$$

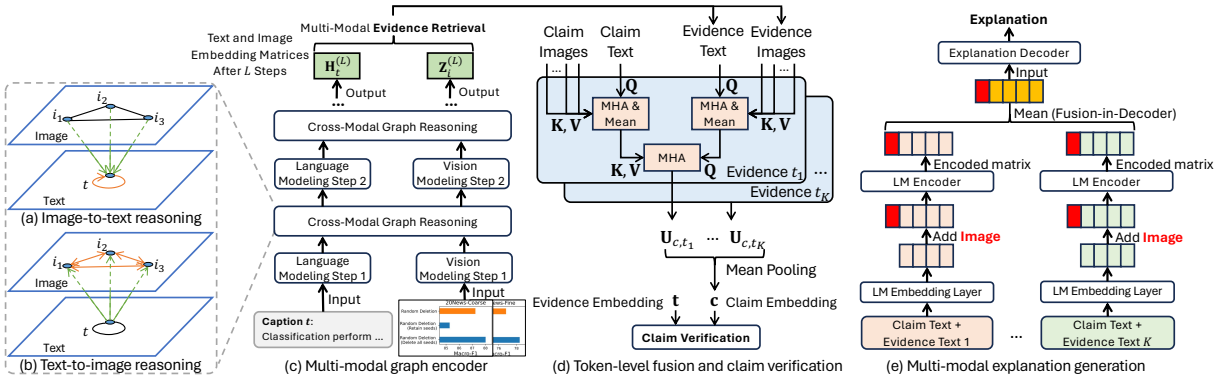


Figure 2: Model architecture. (a-b) Cross-modal graph reasoning. (c) A nested architecture with multi-modal graph reasoning. (d) Multi-modal token-level fusion. (e) Multi-modal explanation generation with Fusion-in-Decoder.

where $i \in \mathcal{I}(t)$ is an evidence image that text t is associated with, $[\cdot||\cdot]$ is concatenation, and $\mathbf{b}_{\text{i2t}} \in \mathbb{R}^{2d}$. We aggregate images by $\hat{\mathbf{z}}_t^{(l)} = \sum_{i \in \mathcal{I}(t)} a_{t,i} \tilde{\mathbf{z}}_{i,\text{CLS}}^{(l)}$. Summarizing above GNN module, we have

$$\hat{\mathbf{z}}_t^{(l)} = f_{\text{GNN}}\Big(\mathbf{h}_{t,\text{CLS}}^{(l)}, \{\mathbf{z}_{i,\text{CLS}}^{(l)} | i \in \mathcal{I}(t)\}; \mathbf{W}_{\text{txt}}, \mathbf{W}_{\text{img}}, \mathbf{b}_{\text{i2t}}\Big). \tag{3}$$

Since textual layer has self-loop, we set $\hat{\mathbf{h}}_t^{(l)} = \tilde{\mathbf{h}}_{t,\text{CLS}}^{(l)}$ without GNN (orange arrow in Fig. 2(a)).

Finally, we integrate the aggregated image and text embeddings into evidence text for cross-modal reasoning. We consider $\hat{\mathbf{z}}_t^{(l)}$ and $\hat{\mathbf{h}}_t^{(l)}$ as *virtual tokens* and concatenate with t's embedding matrix by $\widehat{\mathbf{H}}_t^{(l)} = [\hat{\mathbf{z}}_t^{(l)} || \hat{\mathbf{h}}_t^{(l)} || \mathbf{H}_t^{(l)}]$. To fully unify multi-modal data, we input $\widehat{\mathbf{H}}_t^{(l)}$ to the $(l+1)$-th Transformer step with multi-head attention.

$$\text{MHA}(\mathbf{Q} = \mathbf{W}_Q^{(l)} \mathbf{H}_t^{(l)}, \mathbf{K} = \mathbf{W}_K^{(l)} \widehat{\mathbf{H}}_t^{(l)}, \mathbf{V} = \mathbf{W}_V^{(l)} \widehat{\mathbf{H}}_t^{(l)}). \tag{4}$$

Key $\mathbf{K}$ and value $\mathbf{V}$ are augmented with multi-modal virtual tokens. After multi-layer perceptron and layer normalization, we have evidence text t's embedding matrix $\mathbf{H}_t^{(l+1)}$ at the $(l+1)$-th step.

Text-to-image reasoning. We symmetrically introduce text-to-image reasoning and propagate text embeddings to visual layer for ViT modeling in Fig. 2(b). Since each image i is usually associated with one evidence text t, such as a chart with its caption, text-to-image reasoning simply becomes $\hat{\mathbf{h}}_t^{(l)} = \tilde{\mathbf{h}}_{t,\text{CLS}}^{(l)} = \mathbf{W}_{\text{txt}} \mathbf{h}_{t,\text{CLS}}^{(l)}$. For intra-layer multi-image reasoning, we have GNN module.

$$\hat{\mathbf{z}}_i^{(l)} = f_{\text{GNN}}\Big(\mathbf{z}_{i,\text{CLS}}^{(l)}, \{\mathbf{z}_{i',\text{CLS}}^{(l)} | i' \in \mathcal{I}(t)\}; \mathbf{W}_{\text{img}}, \mathbf{b}_{\text{i2i}}\Big). \tag{5}$$

We consider $\hat{\mathbf{h}}_t^{(l)}$ and $\hat{\mathbf{z}}_i^{(l)}$ as virtual tokens and concatenate with $\mathbf{Z}_i^{(l)}$, i.e., $\widehat{\mathbf{Z}}_i^{(l)} = [\hat{\mathbf{z}}_i^{(l)} || \hat{\mathbf{h}}_t^{(l)} || \mathbf{Z}_i^{(l)}]$. This matrix is input to the $(l+1)$-th ViT step (Dosovitskiy et al., 2021), similarly to Eq. 4.

We repeat image-to-text and text-to-image reasoning inside each Transformer and ViT step for L steps, and obtain a nested encoder in Fig. 2(c).

$$\mathbf{H}_t^{(L)}, \{\mathbf{Z}_i^{(L)}\}_{i \in \mathcal{I}(t)} = f_{\text{Enc}}\big(t, \{i | i \in \mathcal{I}(t)\}\big). \tag{6}$$

We take $\mathbf{h}_t = \mathbf{h}_{t,\text{CLS}}^{(L)}$ as evidence embedding, since it already absorbs both text and image information. For claim c with its associated images, we input to the same multi-modal encoder and obtain claim embedding $\mathbf{h}_c = \mathbf{h}_{c,\text{CLS}}^{(L)}$. We use below contrastive loss as retrieval objective. t' is a negative evidence in the same mini-batch B.

$$\mathcal{L}_{\text{Ret}} = -\sum_{c \in \mathcal{C}_{\text{train}}} \log \frac{\exp(\mathbf{h}_c^\top \mathbf{h}_t)}{\exp(\mathbf{h}_c^\top \mathbf{h}_t) + \sum_{t' \in B \setminus t} \exp(\mathbf{h}_c^\top \mathbf{h}_{t'})}. \tag{7}$$

3.3 Claim Verification with Token- and Evidence-Level Fusion

After we obtain the retrieved evidence with texts and associated images, here we present claim verification using multi-modal evidence. Above multi-modal graph encoder outputs text $\mathbf{H}_t = \mathbf{H}_t^{(L)} \in \mathbb{R}^{P_{\text{txt}} \times d}$ and image embedding matrices $\mathbf{Z}_i = \mathbf{Z}_i^{(L)} \in \mathbb{R}^{P_{\text{img}} \times d}$ for both claims and evidence. P_{txt} and P_{img} respectively denotes the number of textual tokens and visual patches. We allow claims and evidence to interact with each other for claim verification.

Token-level multi-modal fusion. We aim to design an interactive fusion to exchange information between claims and evidence. For each claim c, we have K retrieved evidence texts $\{t_k\}_{k=1}^K$ and images $\mathcal{I}(t_k)$ for each text t_k. We design an interactive fusion method between claims and evidence

at token level, shown by Fig. 2(d). For evidence text t_k and each of its associated images $i \in \mathcal{I}(t_k)$, we leverage a multi-modal multi-head attention for token-level fusion and obtain aggregated image embedding matrix $\mathbf{Z}_{t_k,i} \in \mathbb{R}^{P_{\text{txt}} \times d}$.

$$\mathbf{Z}_{t_k,i} = \text{MHA}(\mathbf{W}_Q \mathbf{H}_{t_k}, \mathbf{W}_K \mathbf{Z}_i, \mathbf{W}_V \mathbf{Z}_i). \quad (8)$$

Since an evidence text t_k has multiple images, we obtain a set $\{\mathbf{Z}_{t_k,i}\}_{i \in \mathcal{I}(t_k)}$. We take mean pooling to obtain a single image embedding matrix by

$$\mathbf{Z}_{t_k} = \text{mean}\big(\{\mathbf{Z}_{t_k,i} | i \in \mathcal{I}(t_k)\}\big). \quad (9)$$

Finally, we unify both text embedding matrix $\mathbf{H}_{t_k}$ and image embedding matrix $\mathbf{Z}_{t_k}$ by

$$\mathbf{U}_{t_k} = \mathbf{W}_1[\mathbf{H}_{t_k} || \mathbf{Z}_{t_k}]. \quad (10)$$

$\mathbf{W}_1 \in \mathbb{R}^{d \times 2d}$, and $\mathbf{U}_{t_k} \in \mathbb{R}^{P_{\text{txt}} \times d}$ is the unified multi-modal matrix for evidence t_k. For claim c and its images $\mathcal{I}(c)$, we repeat Eqs. 8–10 to obtain unified multi-modal matrix for claim $\mathbf{U}_c \in \mathbb{R}^{P_{\text{txt}} \times d}$.

To allow claim-evidence interaction, we have

$$\mathbf{U}_{t_k,c} = \text{MHA}(\mathbf{W}_Q \mathbf{U}_{t_k}, \mathbf{W}_K \mathbf{U}_c, \mathbf{W}_V \mathbf{U}_c). \quad (11)$$

Here $\mathbf{U}_{t_k,c}$ integrates claim c and evidence t_k as well as their associated images. Since a claim has multiple retrieved evidence, we take mean pooling and obtain a single claim embedding matrix $\mathbf{U}_c = \text{mean}\big(\{\mathbf{U}_{t_k,c} | t_k \in \mathcal{T}(c)\}\big)$ where $\mathcal{T}(c)$ is a set of retrieved evidence. Finally, we take [CLS] in $\mathbf{U}_c$ as claim c's embedding $\mathbf{c} = \mathbf{u}_{c,\text{CLS}}$, which fuses claim and evidence at token level. See Fig. 2(d).

Evidence-level hierarchical fusion. We now present evidence embedding. A claim has K retrieved evidence texts, and each text is associated with $|\mathcal{I}(t_k)|$ images, resulting in a hierarchical structure. We propose a hierarchical fusion to obtain multi-modal evidence embedding. Specifically, we have multi-modal text $\mathbf{H}_t = \mathbf{H}_t^{(L)}$ and image embeddings $\mathbf{Z}_i = \mathbf{Z}_i^{(L)}$, output from multi-modal graph encoder in Eq. 6. For each evidence text $t_k \in \mathcal{T}(c)$, we use GNN to aggregate its images.

$$\hat{\mathbf{z}}_{t_k} = f_{\text{GNN}}\Big(\mathbf{h}_{t_k,\text{CLS}}, \{\mathbf{z}_{i,\text{CLS}} | i \in \mathcal{I}(t_k)\}; \mathbf{W}_{\text{txt}}, \mathbf{W}_{\text{img}}, \mathbf{b}_{\text{i2t}}\Big). \quad (12)$$

The aggregated image embedding is then combined with evidence text embedding by $\mathbf{t}_k = \mathbf{W}_2[\mathbf{h}_{t_k,\text{CLS}} || \hat{\mathbf{z}}_{t_k}]$ where $\mathbf{W}_2 \in \mathbb{R}^{d \times 2d}$ and $\mathbf{t}_k$ is evidence embedding. Finally, we use claim embedding $\mathbf{c}$ as query to aggregate evidence embeddings.

$$\mathbf{t} = f_{\text{GNN}}\Big(\mathbf{c}, \{\mathbf{t}_k | t_k \in \mathcal{T}(c)\}; \mathbf{W}_{\text{txt}}, \mathbf{W}_{\text{evid}}, \mathbf{b}_{\text{e2c}}\Big). \quad (13)$$

Having obtained claim and evidence embeddings, we input them into a classifier for verification, $\hat{\mathbf{y}} = \text{softmax}(f_{\text{MLP}}([\mathbf{c}||\mathbf{t}]))$ with below loss.

$$\mathcal{L}_{\text{Ver}} = -\sum_{y' \in \mathcal{Y}} y' \log \hat{y}'. \quad (14)$$

3.4 Multi-Modal Explanation Generation

We use multi-modal data to generate explanations to justify the ruling process. We design Fusion-in-Decoder (Fig. 2(e)) and consistency regularizer.

Multi-modal Fusion-in-Decoder. We are given a claim text c and each of its evidence texts $t_k \in \mathcal{T}(c)$, we concatenate their raw texts into a single textual sequence with [SEP] token as separator, i.e., $[c||[SEP]||t_k]$. In language models, there is an embedding look-up table before language encoder. Textual sequence is first mapped to token embeddings in this look-up table, which are then summed up with positional encodings. We thus obtain embedding matrix $\mathbf{E}_{c,t_k}$ for the concatenated sequence using this look-up table. We then concatenate the associated image embeddings of claim and evidence text to obtain a multi-modal matrix.

$$\widehat{\mathbf{E}}_{c,t_k} = [\underbrace{\mathbf{W}_{\text{img}}\mathbf{z}_{c,\text{CLS}}||...}_{\text{claim's images}} || \underbrace{\mathbf{W}_{\text{img}}\mathbf{z}_{t_k,\text{CLS}}||...}_{\text{evidence text's images}} ||\mathbf{E}_{c,t_k}] \quad (15)$$

$\mathbf{z}_{c,\text{CLS}}$ and $\mathbf{z}_{t_k,\text{CLS}}$ are respectively claim c's and evidence text t_k's image embeddings, obtained from multi-modal graph encoder in Eq. 6. We project them using $\mathbf{W}_{\text{img}}$ to the same embedding space as texts. We input the concatenated multi-modal embedding matrix $\widehat{\mathbf{E}}_{c,t_k}$ to language encoder, T5 (Raffel et al., 2020), and obtain $\tilde{\mathbf{E}}_{c,t_k} = f_{\text{LMEnc}}(\widehat{\mathbf{E}}_{c,t_k})$.

Since a claim c has multiple retrieved evidence $\mathcal{T}(c)$, we design multi-modal Fusion-in-Decoder module to capture all evidence for explanation generation, i.e., $\hat{e} = f_{\text{LMDec}}(\bar{\mathbf{E}}_{c,t})$ where $\bar{\mathbf{E}}_c = \text{mean}(\tilde{\mathbf{E}}_{c,t_1}, ..., \tilde{\mathbf{E}}_{c,t_K})$. This module is illustrated by Fig. 2(e). We have generation loss below.

$$\mathcal{L}_{\text{Exp}} = -\sum_{e_j} \log p(e_j | e_{<j}, \bar{\mathbf{E}}_c) \quad (16)$$

$e_{<j}$ denotes the tokens generated prior to e_j, and $p(e_j|\cdot) = \text{softmax}(\mathscr{L}(e_j|\cdot))$ is the probability of e_j, calculated by normalizing logits $\mathscr{L}(e_j|\cdot)$.

Consistency regularizer. The generated explanation should consistently justify the predicted verification. Thus, we design a consistency regularizer to achieve this goal. Specifically, we do mean pooling for logits of all the tokens in explanation by

$$\mathscr{L} = \text{mean}(\mathscr{L}(e_1|\cdot), ..., \mathscr{L}(e_j|\cdot), ...). \quad (17)$$

$\mathscr{L} \in \mathbb{R}^V$ is a V-dimensional embedding where V is the length of language model vocabulary. $\mathscr{L}$ contains the information of explanation, and its embedding should also reveal verification label. Thus, we input $\mathscr{L}$ to a classifier to predict verification label by $\hat{\mathbf{y}}_e = \text{softmax}(f_{\text{MLP}}(\mathscr{L}))$. Finally, we minimize the difference between the predicted label $\hat{\mathbf{y}}$ of the verification module and $\hat{\mathbf{y}}_e$ using

$$\mathcal{L}_{\text{Reg}} = \text{KL}(\hat{\mathbf{y}}||\hat{\mathbf{y}}_e) + \text{KL}(\hat{\mathbf{y}}_e||\hat{\mathbf{y}}) - \sum_{y' \in \mathcal{Y}} y' \log \hat{y}'_e. \tag{18}$$

We sum up two KL divergences (Bishop and Nasrabadi, 2006) to remove its asymmetry. We also add a cross-entropy loss for the predicted label of logits. The overall loss function becomes

$$\mathcal{L} = \mathcal{L}_{\text{Ver}} + \mathcal{L}_{\text{Exp}} + \lambda \mathcal{L}_{\text{Reg}}. \tag{19}$$

$\lambda = 0.5$ is a hyperparameter for the importance of regularizer. We summarize the learning in Algos. 1–2. See Appendix C for complexity analysis.

4 Dataset Creation

We create a scientific dataset. We present main creation here and put more details in Appendix E.

Data source. We use AI papers as source. Other domains are future work. We have 5 categories, and each category has 3 conferences, AI (AAAI, IJCAI, UAI), ML (NeurIPS, ICML, ICLR), NLP (ACL, EMNLP, NAACL), CV (CVPR, ICCV, ECCV), Data Mining (KDD, WWW, WSDM), i.e., totally 15 conferences. For each conference, we collect recent 4 proceedings. For each proceeding, we select 5 papers with chart, with totally 300 papers.

Within each paper, we collect its chart with caption as multi-modal evidence. The sentences in the main text that mention the chart usually contain scientific discoveries or claims. However, not every claim is checkworthy and not every chart is clear. To ensure the quality, our annotators (one postdoc and three PhD students specializing in AI) filter out inappropriate claims and charts and search for other papers with high-quality data. Finally, we have 300 claims with corresponding 300 charts and captions. These charts include line, bar, pie, scatterplot, and heat map. See Table 6 for details.

Data augmentation. Usually, the charts support the claims, thus above 300 claims, undergoing manual double check by annotators, are labeled as "SUPPORT". To create claims refuted by the charts, we follow Wadden et al. (2020) and ask annotators to write negation for the 300 claims, taking precautions not to bias the negation by using obvious keywords, like "not". We finally obtain 600 claims, half supported and half refuted by the charts.

To augment the dataset, we follow Schlichtkrull et al. (2024) and use GPT-4o (Achiam et al., 2023) to generate 600 more claims. For each of 300 charts with captions, we use the prompt in Appendix E to generate two more claims, one supported and one refuted by the chart. After generation, we have 1,200 claims, i.e., 600 natural and 600 generated claims. Our annotators rigorously check the generation and make manual corrections when necessary to ensure the dataset is correct and high-quality.

Explanation. A significant portion of papers do not have explanations for claims. Moreover, we do not have explanations for the generated claims. For consistency, we use GPT-4o to generate explanations with prompt in Appendix E. The generation is rigorously checked and corrected by annotators if there is erroneous or unclear description.

Size. Finally, we have 1,200 claims with explanations and 300 charts with captions. The size is comparable to textual scientific datasets, e.g., 1,409 claims in SciFact (Wadden et al., 2020), 1,448 claims in BearFact (Wuehrl et al., 2024), 1,504 claims in Check-COVID (Wang et al., 2023). There are two constraints for the size. For one, the number of AI papers with checkworthy claims and readable charts is limited. Our annotators have tried hard to obtain appropriate papers. For the other, the creation requires experts to manually analyze the charts. The number of domain experts is limited. We consider creating larger-size dataset as future work. Also, we follow Akhtar et al. (2024) and do not include NEI label for clarity and consistency. We can also add NEI by simply replacing the gold evidence of each claim with random evidence.

Table 2: Dataset statistics.

Name	#Claims	#Evidence Texts	#Images	Explanation
AIChartClaim	1,200	300	300	Yes
ChartCheck	10,038	1,615	1,615	Yes
Mocheg	15,601	33,880	13,052	Yes
MR2	13,785	91,347	105,132	No

5 Experiments

Datasets. We present main statistics of the datasets in Table 2. Besides **AIChartClaim**, we have **ChartCheck** (Akhtar et al., 2024), a general-domain chart dataset. **Mocheg** (Yao et al., 2023) is

Table 3: Result of evidence retrieval (%). Standard deviation of BM25 is 0, because it is a deterministic approach.

Model	AIChartClaim			ChartCheck			Mocheg			MR2		
	MAP	Prec@3	Rec@3	MAP	Prec@3	Rec@3	MAP	Prec@3	Rec@3	MAP	Prec@3	Rec@3
BM25	49.5±0.0	17.6±0.0	52.9±0.0	40.6±0.0	14.6±0.0	43.6±0.0	27.3±0.0	21.2±0.0	26.8±0.0	11.7±0.0	17.5±0.0	8.4±0.0
RAV	59.0±0.8	21.6±0.3	64.7±1.0	59.9±0.2	21.6±0.2	64.8±0.7	39.0±0.8	29.7±0.6	37.6±0.8	15.2±0.3	22.1±0.6	10.7±0.4
JustiLM	62.0±1.2	22.4±1.0	64.2±3.0	58.3±0.5	21.1±0.2	63.4±0.7	39.5±0.4	29.8±0.2	38.0±0.2	14.5±0.2	20.7±0.5	10.1±0.2
MochegModel	65.9±0.0	23.3±0.0	70.0±0.0	58.1±0.0	21.0±0.0	63.0±0.1	36.0±0.0	27.9±0.0	38.2±0.0	16.7±0.0	23.2±0.0	11.2±0.0
TransXH+ViT	62.0±2.5	22.7±0.7	68.1±2.1	60.8±0.3	21.6±0.1	64.8±0.3	39.8±0.5	30.2±0.3	38.1±0.4	17.8±0.6	28.2±0.3	12.6±0.5
MEVER w/o images	69.8±1.0	24.9±0.5	74.7±1.5	58.3±0.9	21.2±0.3	63.7±0.9	38.6±2.9	29.3±2.5	37.3±2.9	15.0±0.2	21.7±0.4	10.6±0.2
MEVER (ours)	**71.4±0.3**	**25.4±0.1**	**76.1±0.2**	**63.6±0.3**	**22.7±0.1**	**68.0±0.3**	**41.6±0.4**	**31.7±0.3**	**40.1±0.2**	**19.5±0.4**	**29.8±0.4**	**13.1±0.3**

Table 4: Claim verification with *Macro F1* score (%).

Model	AIChartClaim		ChartCheck		Mocheg		MR2	
	Gold	Retrieved	Gold	Retrieved	Gold	Retrieved	Gold	Retrieved
KGAT	68.7±0.4	68.8±0.3	62.4±0.2	60.4±0.3	47.4±0.6	45.1±0.5	65.0±1.5	64.3±1.1
CausalWalk	68.4±2.3	69.0±2.4	60.4±2.8	61.3±0.9	38.6±0.7	30.1±0.6	64.6±2.5	65.3±4.8
CORRECT	67.9±1.5	70.0±1.2	62.3±0.4	61.5±0.3	46.0±0.9	45.9±1.7	70.0±1.5	70.4±3.0
TransformerXH	68.6±0.2	69.1±0.6	62.9±0.4	62.0±0.9	43.8±0.9	44.3±0.6	68.1±1.2	70.9±1.7
TransformerXH++	69.4±0.9	68.3±1.0	63.2±0.2	62.6±0.6	45.0±1.4	44.5±0.1	69.6±0.4	72.5±1.0
MochegModel	57.9±1.7	59.0±2.4	56.7±1.6	57.8±1.3	45.6±1.4	45.5±1.2	64.8±0.3	68.0±0.4
MR2Retrieved	69.9±0.5	69.8±0.2	62.7±0.1	62.3±0.9	44.0±0.2	44.5±0.6	73.9±1.2	71.2±0.8
CCN	69.8±0.9	69.3±0.2	62.6±0.4	62.6±0.5	44.9±0.4	47.5±0.4	73.5±1.3	75.5±2.1
ESCNet	69.5±0.6	69.6±0.4	60.7±0.5	61.3±0.5	46.4±0.3	47.4±0.5	73.5±0.6	75.0±0.9
ECENet	70.0±0.7	70.2±0.7	60.9±0.8	61.3±0.9	45.5±1.2	46.7±1.8	72.4±1.2	74.2±0.8
MultiKE-GAT	66.6±0.4	67.3±0.5	60.5±0.5	60.6±0.8	39.9±1.4	46.2±1.0	67.0±1.8	71.6±2.0
GPT-4o	51.0±1.7	41.7±1.2	43.9±0.8	49.8±1.8	**48.7±3.7**	44.5±3.5	68.7±2.4	63.7±2.8
ChartBERT	42.6±4.7	43.0±1.6	55.7±0.6	40.9±2.9	N.A.	N.A.	N.A.	N.A.
UniChart	69.3±1.6	68.4±0.4	62.6±0.6	62.3±0.1	N.A.	N.A.	N.A.	N.A.
ChartGemma	69.1±0.4	68.6±1.4	63.3±0.3	63.9±0.3	N.A.	N.A.	N.A.	N.A.
JustiLM	65.4±0.7	65.3±0.2	61.3±0.5	62.3±0.1	38.5±0.7	44.4±0.4	69.7±1.1	72.1±0.5
RAV	67.5±0.4	67.6±0.2	61.8±1.2	60.3±1.5	45.8±1.1	42.5±0.4	65.0±1.7	61.9±2.6
MEVER w/o images	66.3±0.6	68.5±0.6	61.6±1.8	61.9±0.7	45.4±0.7	46.2±0.4	67.5±2.9	70.3±1.1
MEVER (ours)	**71.6±0.7**	**71.6±0.4**	**64.3±0.6**	**64.1±0.3**	**48.3±2.1**	**49.7±1.2**	**76.0±0.7**	**77.7±0.1**

multi-modal dataset with explanation. **MR2** (Hu et al., 2023) has images for claims and evidence, but no explanations. Appendix E has more details.

Implementation. We set L to 12 and d to 768. We initialize the model with scientific parameters (Beltagy et al., 2019) for AIChartClaim and with general parameters (Devlin et al., 2019) for others. Each result is obtained by 3 independent runs. Experiments are done on 4 NVIDIA A100 GPUs. More details are in Appendix F.

5.1 Evidence Retrieval

Baselines. ***i*) Text-only retrieval**, BM25 (Robertson et al., 2009), RAV (Zheng et al., 2024), JustiLM (Zeng and Gao, 2024). ***ii*) Multi-modal retrieval**, MochegModel (Yao et al., 2023). Since our retriever is built on TransformerXH and ViT, we add one multi-modal baseline, TransXH+ViT. We add an ablation by removing images from our model.

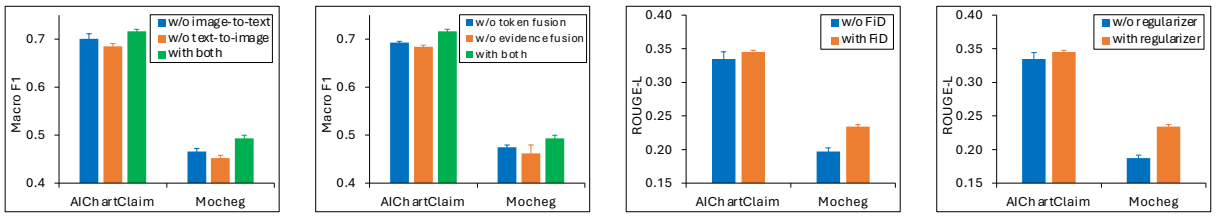
Analysis. If the dataset has training-test split, we follow its split. Otherwise, we split 80% for training, among which 10% for validation. We follow Yao et al. (2023) and report MAP, Precision@κ, and Recall@κ in Table 3 with $\kappa = 3$. We further vary κ in $\{1, 3, 5, 7\}$ in Appendix G.1. Multi-modal retrieval outperforms text-only methods, since images bring useful information. Our model performs better than them, since our multi-modal graph encoder uses a nested architecture to integrate multi-modal data. The ablated model drops the result, verifying that images provide useful information.

5.2 Claim Verification

Baselines. ***i*) Text-only**, KGAT (Liu et al., 2020), CausalWalk (Zhang et al., 2024a), CORRECT (Zhang and Lee, 2025), TransformerXH (Zhao et al., 2020). ***ii*) Multi-modal**, MochegModel (Yao et al., 2023), MR2Retrieved (Hu et al., 2023), CCN (Abdelnabi et al., 2022), ESCNet (Zhang et al., 2024b), ECENet (Zhang et al., 2023b), MultiKE-GAT (Cao et al., 2024), GPT-4o (Achiam et al., 2023). ***iii*) Chart-based**, ChartBERT (Akhtar et al., 2023), UniChart (Masry et al., 2023), ChartGemma (Masry et al., 2025). ***iv*) Retrieval-augmented**,

Table 5: Explanation generation with *ROUGE-L*, *METEOR*, and *BLEU-2* scores (%) with *retrieved evidence setting*. See Appendix G.3 for results on gold evidence setting. MR2 dataset does not have explanations.

Model	AIChartClaim			ChartCheck			Mocheg		
	ROUGE-L	METEOR	BLEU-2	ROUGE-L	METEOR	BLEU-2	ROUGE-L	METEOR	BLEU-2
JustiLM	21.7±0.7	16.8±1.0	12.4±1.0	34.7±1.8	30.6±1.9	23.5±1.9	19.3±0.5	17.7±0.3	12.1±0.7
MochegModel	33.4±1.0	25.7±1.4	18.2±1.5	39.6±0.3	33.8±0.4	24.5±0.5	18.7±0.4	18.7±0.3	14.8±0.1
ECENet	32.3±0.7	26.8±0.6	19.2±1.1	39.7±0.6	34.7±0.5	25.7±0.8	20.5±0.6	15.4±0.1	12.2±0.2
DePlot+FlanT5	33.5±1.1	25.8±1.4	18.3±1.5	39.9±0.1	34.0±0.2	24.7±0.2	19.7±0.5	19.0±0.5	16.2±0.7
GPT-4o	18.2±0.6	23.2±0.5	16.1±0.3	17.7±0.4	27.4±0.2	16.2±0.3	18.2±1.6	18.8±1.8	10.9±0.8
UniChart	33.2±0.2	25.9±0.4	19.6±0.1	40.1±0.2	34.6±0.2	26.6±0.2	N.A.	N.A.	N.A.
ChartGemma	33.5±0.3	27.1±0.2	20.4±0.1	40.1±0.1	35.7±0.2	26.4±0.1	N.A.	N.A.	N.A.
MEVER w/o images	33.7±0.1	26.3±0.6	21.1±0.1	39.7±0.1	34.8±0.3	25.5±0.5	22.9±0.1	19.0±0.1	15.1±0.1
MEVER (ours)	**34.5±0.2**	**27.8±0.4**	**21.3±0.4**	**40.8±0.1**	**36.2±0.2**	**27.2±0.4**	**23.4±0.3**	**20.0±0.3**	**16.3±0.3**



(a) Effect of cross-modal reasoning (b) Effect of token and evidence fusion (c) Effect of multi-modal FiD (d) Effect of consistency regularizer

Figure 3: Model analysis on AIChartClaim and Mocheg datasets.

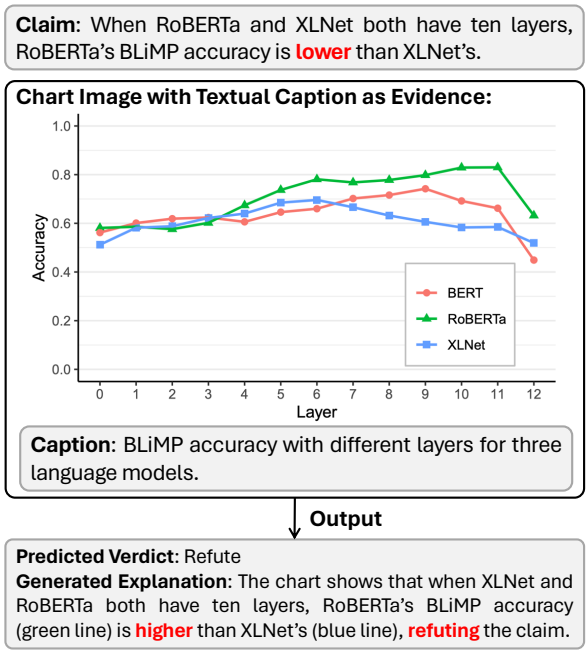


Figure 4: Case study on AIChartClaim dataset.

JustiLM (Zeng and Gao, 2024), RAV (Zheng et al., 2024). We convert TransformerXH to its multi-modal version, TransformerXH++. We add our ablated model by removing images. For instruction-tuned models, we use 5-shot setting. Since chart-based models are specifically designed for charts, we do not report their result on Mocheg and MR2.

Gold v.s. retrieved setting. For gold setting, we have gold multi-modal evidence. For retrieved setting, we first retrieve multi-modal evidence, which is then used for verification. Since our retriever achieves the best result, we use our retrieved evidence for our model and all baselines for fairness.

Analysis. As in Wadden et al. (2022), we show Macro F1 in Table 4. See Appendix G.2 for Micro F1. Multi-modal baselines tend to outperform others, showing the strength of images. However, all of them model evidence-level fusion. Our model is better than them, showing the strength of both token- and evidence-level fusion. Though evidence retrieval on Mocheg and MR2 is more difficult, MEVER still outperforms baselines on retrieved setting, showing that MEVER is not affected by the difficulty of evidence retrieval. GPT-4o does not perform well on AIChartClaim, since our dataset creation inputs gold label to GPT-4o to generate explanations, but the current verification does not input gold labels, leading to inaccurate result.

See Table 7 for more results on AIChartClaim.

5.3 Explanation Generation

Baselines. ***i*) Text-only**, JustiLM (Zeng and Gao, 2024), a retrieval-augmented baseline. ***ii*) Multi-modal**, MochegModel (Yao et al., 2023), ECENet (Zhang et al., 2023b), DePlot+FlanT5 (Akhtar et al., 2024), GPT-4o (Achiam et al., 2023). ***iii*) Chart-based**, UniChart (Masry et al., 2023), ChartGemma (Masry et al., 2025). We add our ablated model.

Analysis. We follow Yao et al. (2023) and show

ROUGE-L, METEOR, and BLEU-2 with retrieved setting in Table 5. See Appendix G.3 for ROUGE-1, ROUGE-2, BLEU-4, and VLM-as-a-Judge. Our model outperforms baselines, since consistency regularizer uses predicted label to regularize the explanation towards accurate generation. GPT-4o does not perform well, because we use its predicted labels in Sec. 5.2 as input to generate explanation to keep consistent with our model. Its inaccurate predicted verdicts influence explanation generation.

5.4 Model Analysis

Cross-modal reasoning. We respectively remove image-to-text and text-to-image reasoning from the model in Fig. 3(a). The complete model performs the best, showing the strength of both reasoning. Removing text-to-image leads to the worst result, since images do not contain enough information.

Token and evidence fusion. We remove each fusion in Fig. 3(b). The complete model outperforms others, showing the strength of both fusions. Removing evidence fusion hurts the result, since claim embedding, though with evidence information after claim-evidence interaction, still needs evidence explicitly for accurate verification.

Multi-modal Fusion-in-Decoder (FiD). Fig. 3(c) shows that removing FiD hurts explanation generation quality, since FiD aggregates multiple evidence for generation, and disregarding it leads to insufficient multi-evidence reasoning.

Consistency regularizer. Fig. 3(d) shows that consistency regularizer controls the generation towards the predicted label for consistent explanation, thus improving explanation quality.

Case study. In Fig. 4, our model reasons over text and image to correctly verify the claim, and the explanation clearly justifies the prediction. This visualization shows the effectiveness of our model. For more case studies, our submitted code produces predicted labels and explanations for all claims.

5.5 Analysis of the AIChartClaim Dataset

Chart type. We collect 300 charts in total and provide chart types in Table 6. Since line charts and bar charts are the most commonly used charts in academic papers to express scientific results and discoveries, they are the top-2 chart types in the dataset. Here "Bar with Number" indicates a bar chart where authors also put numerical result on top of each bar, same for "Line with Number". We differentiate them from bar and line charts, because usually models need to further recognize the numerical value on the bar and line to verify the claims.

Table 6: AIChartClaim dataset statistics.

Chart Type	Number of Charts
Line	203
Bar	61
Bar with Number	16
Line with Number	6
Line + Bar	5
Scatterplot	4
Scatterplot with Lines	3
Others (Pie and Heat Map)	2
Total	300

Table 7: Claim verification of our model MEVER on test set of each chart type (%).

Chart Type	AIChartClaim	
	Micro F1	Macro F1
Line	75.6	75.6
Bar	67.3	67.0
Bar with Number	56.3	56.3
Line with Number	75.0	75.3
Line + Bar	62.5	61.9
Others (Scatterplot, Pie, Heat Map)	55.6	55.6

Analysis on each chart type. We further report claim verification result of our model on each chart type in Table 7. Overall, our model performs the best on line charts, including "Line" and "Line with Number". One reason is that our dataset has the most number of line charts, which contain more information than other chart types to train the model. The performance on bar charts, including "Bar", "Bar with Number", and "Line + Bar", is also decent. Our model does not perform well on other chart types, potentially due to insufficient training data. However, we emphasize that line charts and bar charts are the most common charts in AI papers, and it is quite difficult to obtain sufficient scatterplots, pie charts, and heat maps from AI papers.

6 Conclusion

We propose MEVER with evidence retrieval, claim verification, and explanation. We create a scientific dataset in AI domain. A future work is to explore multi-modal knowledge graph for verification.

Acknowledgments

This work was in part supported by NSF awards #1934782 and #2114824. Some of the research results were obtained using computational resources provided by NAIRR award #240336.

Limitations

Here we identify two limitations in terms of evidence type and our proposed AIChartClaim dataset.

Evidence type. Our model is proposed mainly in the multi-modal setting where we assume both texts and images are available. If images are absent from the dataset and we have texts only, we may need additional effort to obtain images, so that our model can use both modalities for claim verification. As suggested by Abdelnabi et al. (2022), one potential solution is to search for relevant images given evidence texts, such as collecting charts using the captions or paper titles. Since our paper does not study such cross-modal information search, we leave it as a future work.

AIChartClaim dataset. ***i*) Size.** The size of our created dataset is already comparable to other scientific text-only datasets. The size of our dataset is limited by two main factors. First, the number of AI papers with checkworthy claims and clearly readable charts is limited. We have tried our best to obtain appropriate AI papers. Second, the dataset creation process requires domain experts with scientific knowledge to check and analyze the charts, thus the creation is quite effort-consuming and knowledge-dependent. The number of domain experts is also limited. We consider creating a large-size dataset as a future work.

***ii*) Domain.** We are AI researchers and mainly focus on AI domain for now. Since there is not an existing scientific dataset with multi-modal data and explanation, our dataset is the first step to advance the research community. We are interested in collecting datasets in other domains, such as biomedicine, in the future.

***iii*) Verdict.** Following existing chart datasets, e.g., ChartCheck (Akhtar et al., 2024), our dataset has two types of verdicts, SUPPORT and REFUTE. Some other datasets have the third verdict, NEI for Not Enough Info. Following Wang et al. (2023), we can also introduce the third verdict by simply replacing the gold evidence of claims with random evidence. But for clarity and consistency purpose, we do not introduce NEI verdict in our dataset.

Ethics Statement

We do not foresee any undesired implications stemming from our work. Conversely, we hope that our work can advance AI Ethics research.

References

Sahar Abdelnabi, Rakibul Hasan, and Mario Fritz. 2022. Open-domain, content-based, multi-modal fact-checking of out-of-context images via online resources. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14940–14949.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Mubashara Akhtar, Oana Cocarascu, and Elena Simperl. 2023. Reading and reasoning over chart images for evidence-based automated fact-checking. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 399–414.

Mubashara Akhtar, Nikesh Subedi, Vivek Gupta, Sahar Tahmasebi, Oana Cocarascu, and Elena Simperl. 2024. Chartcheck: Explainable fact-checking over real-world chart images. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 13921–13937.

Rami Aly, Zhijiang Guo, Michael Schlichtkrull, James Thorne, Andreas Vlachos, Christos Christodoulopoulos, Oana Cocarascu, and Arpit Mittal. 2021. Feverous: Fact extraction and verification over unstructured and structured information. In *35th Conference on Neural Information Processing Systems, NeurIPS 2021*. Neural Information Processing Systems foundation.

Pepa Atanasova. 2024. Generating fact checking explanations. In *Accountable and Explainable Methods for Complex Reasoning over Text*, pages 83–103. Springer.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. Scibert: A pretrained language model for scientific text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3615–3620.

Iz Beltagy, Matthew E Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*.

Christopher M Bishop and Nasser M Nasrabadi. 2006. *Pattern recognition and machine learning*, volume 4. Springer.

Han Cao, Lingwei Wei, Wei Zhou, and Songlin Hu. 2024. Multi-source knowledge enhanced graph attention networks for multimodal fact verification. In *2024 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE.

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.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*.

Will Hamilton, Zhitao Ying, and Jure Leskovec. 2017. Inductive representation learning on large graphs. *Advances in neural information processing systems*, 30.

Xuming Hu, Zhijiang Guo, Junzhe Chen, Lijie Wen, and Philip S Yu. 2023. Mr2: A benchmark for multimodal retrieval-augmented rumor detection in social media. In *Proceedings of the 46th international ACM SIGIR conference on research and development in information retrieval*, pages 2901–2912.

Kung-Hsiang Huang, Hou Pong Chan, Yi R Fung, Haoyi Qiu, Mingyang Zhou, Shafiq Joty, Shih-Fu Chang, and Heng Ji. 2024. From pixels to insights: A survey on automatic chart understanding in the era of large foundation models. *IEEE Transactions on Knowledge and Data Engineering*.

Gautier Izacard, Patrick Lewis, Maria Lomeli, Lucas Hosseini, Fabio Petroni, Timo Schick, Jane Dwivedi-Yu, Armand Joulin, Sebastian Riedel, and Edouard Grave. 2023. Atlas: Few-shot learning with retrieval augmented language models. *Journal of Machine Learning Research*, 24(251):1–43.

Neema Kotonya and Francesca Toni. 2020a. Explainable automated fact-checking for public health claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7740–7754.

Neema Kotonya and Francesca Toni. 2020b. Explainable automated fact-checking for public health claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7740–7754.

Seongyun Lee, Seungone Kim, Sue Park, Geewook Kim, and Minjoon Seo. 2024. Prometheus-vision: Vision-language model as a judge for fine-grained evaluation. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 11286–11315.

Fangyu Liu, Francesco Piccinno, Syrine Krichene, Chenxi Pang, Kenton Lee, Mandar Joshi, Yasemin Altun, Nigel Collier, and Julian Martin Eisenschlos. 2023. Matcha: Enhancing visual language pretraining with math reasoning and chart derendering. In *The 61st Annual Meeting Of The Association For Computational Linguistics*.

Fuxiao Liu, Xiaoyang Wang, Wenlin Yao, Jianshu Chen, Kaiqiang Song, Sangwoo Cho, Yaser Yacoob, and Dong Yu. 2024a. Mmc: Advancing multimodal chart understanding with large-scale instruction tuning. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1287–1310.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024b. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26296–26306.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. 2024c. Llava-next: Improved reasoning, ocr, and world knowledge.

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

Grace Luo, Trevor Darrell, and Anna Rohrbach. 2021. Newsclippings: Automatic generation of out-of-context multimodal media. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6801–6817.

Haoran Luo, Guanting Chen, Yandan Zheng, Xiaobao Wu, Yikai Guo, Qika Lin, Yu Feng, Zemin Kuang, Meina Song, Yifan Zhu, et al. 2025. Hypergraphrag: Retrieval-augmented generation via hypergraph-structured knowledge representation. *arXiv preprint arXiv:2503.21322*.

Ahmed Masry, Parsa Kavehzadeh, Xuan Long Do, Enamul Hoque, and Shafiq Joty. 2023. Unichart: A universal vision-language pretrained model for chart comprehension and reasoning. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14662–14684.

Ahmed Masry, Megh Thakkar, Aayush Bajaj, Aaryaman Kartha, Enamul Hoque, and Shafiq Joty. 2025. Chartgemma: Visual instruction-tuning for chart reasoning in the wild. In *Proceedings of the 31st International Conference on Computational Linguistics: Industry Track*, pages 625–643.

Fanqing Meng, Wenqi Shao, Quanfeng Lu, Peng Gao, Kaipeng Zhang, Yu Qiao, and Ping Luo. 2024. Chartassistant: A universal chart multimodal language model via chart-to-table pre-training and multitask instruction tuning. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 7775–7803.

Shreyash Mishra, S Suryavardan, Amrit Bhaskar, Parul Chopra, Aishwarya N Reganti, Parth Patwa, Amitava Das, Tanmoy Chakraborty, Amit P Sheth, Asif Ekbal, et al. 2022. Factify: A multi-modal fact verification dataset. In *DE-FACTIFY@ AAAI*.

Cong-Duy Nguyen, Xiaobao Wu, Duc Anh Vu, Shuai Zhao, Thong Nguyen, and Anh Tuan Luu. 2025a. Cutpaste&find: Efficient multimodal hallucination detector with visual-aid knowledge base. *arXiv preprint arXiv:2502.12591*.

Thong Thanh Nguyen, Xiaobao Wu, Yi Bin, Cong-Duy T Nguyen, See-Kiong Ng, and Anh Tuan Luu. 2025b. Motion-aware contrastive learning for temporal panoptic scene graph generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 6218–6226.

Liangming Pan, Xiaobao Wu, Xinyuan Lu, Luu Anh Tuan, William Yang Wang, Min-Yen Kan, and Preslav Nakov. 2023. Fact-checking complex claims with program-guided reasoning. In *Proceedings of the 61st annual meeting of the association for computational linguistics (volume 1: long papers)*, pages 6981–7004.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67.

Stephen Robertson, Hugo Zaragoza, et al. 2009. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends® in Information Retrieval*, 3(4):333–389.

Daniel Russo, Serra Sinem Tekiroğlu, and Marco Guerini. 2023. Benchmarking the generation of fact checking explanations. *Transactions of the Association for Computational Linguistics*, 11:1250–1264.

Michael Schlichtkrull, Zhijiang Guo, and Andreas Vlachos. 2024. Averitec: A dataset for real-world claim verification with evidence from the web. *Advances in Neural Information Processing Systems*, 36.

Shyam Subramanian and Kyumin Lee. 2020. Hierarchical evidence set modeling for automated fact extraction and verification. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7798–7809.

J Thorne, A Vlachos, C Christodoulopoulos, and A Mittal. 2018. Fever: a large-scale dataset for fact extraction and verification. In *Proceedings of the 16th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Sheffield.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. 2018. Graph attention networks. In *International Conference on Learning Representations*.

David Wadden, Shanchuan Lin, Kyle Lo, Lucy Lu Wang, Madeleine van Zuylen, Arman Cohan, and Hannaneh Hajishirzi. 2020. Fact or fiction: Verifying scientific claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7534–7550.

David Wadden, Kyle Lo, Lucy Lu Wang, Arman Cohan, Iz Beltagy, and Hannaneh Hajishirzi. 2022. Multivers: Improving scientific claim verification with weak supervision and full-document context. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 61–76.

Gengyu Wang, Kate Harwood, Lawrence Chillrud, Amith Ananthram, Melanie Subbiah, and Kathleen Mckeown. 2023. Check-covid: Fact-checking covid-19 news claims with scientific evidence. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 14114–14127.

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

Xiaobao Wu, Liangming Pan, William Yang Wang, and Anh Tuan Luu. 2024a. Updating language models with unstructured facts: Towards practical knowledge editing. *CoRR*.

Xiaobao Wu, Liangming Pan, William Yang Wang, and Luu Anh Tuan. 2024b. Akew: Assessing knowledge editing in the wild. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 15118–15133.

Amelie Wuehrl, Yarik Menchaca Resendiz, Lara Grimminger, and Roman Klinger. 2024. What makes medical claims (un) verifiable? analyzing entity and relation properties for fact verification. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2046–2058.

Michiharu Yamashita, Thanh Tran, Delvin Ce Zhang, and Dongwon Lee. 2025. Unmasking fake careers: Detecting machine-generated career trajectories via multi-layer heterogeneous graphs. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 20893–20908.

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

46th International ACM SIGIR Conference on Research and Development in Information Retrieval, pages 2733–2743.

Zhenrui Yue, Huimin Zeng, Lanyu Shang, Yifan Liu, Yang Zhang, and Dong Wang. 2024. Retrieval augmented fact verification by synthesizing contrastive arguments. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10331–10343, Bangkok, Thailand. Association for Computational Linguistics.

Fengzhu Zeng and Wei Gao. 2024. Justilm: Few-shot justification generation for explainable fact-checking of real-world claims. *Transactions of the Association for Computational Linguistics*, 12:334–354.

Ce Zhang and Hady W Lauw. 2020. Topic modeling on document networks with adjacent-encoder. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 6737–6745.

Congzhi Zhang, Linhai Zhang, and Deyu Zhou. 2024a. Causal walk: Debiasing multi-hop fact verification with front-door adjustment. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 19533–19541.

Delvin Ce Zhang and Hady W Lauw. 2023. Topic modeling on document networks with dirichlet optimal transport barycenter. *IEEE Transactions on Knowledge and Data Engineering*, 36(3):1328–1340.

Delvin Ce Zhang and Dongwon Lee. 2025. Correct: Context- and reference-augmented reasoning and prompting for fact-checking. In *Proceedings of 2025 Annual Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics*.

Delvin Ce Zhang, Rex Ying, and Hady W Lauw. 2023a. Hyperbolic graph topic modeling network with continuously updated topic tree. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 3206–3216.

Fanrui Zhang, Jiawei Liu, Jingyi Xie, Qiang Zhang, Yongchao Xu, and Zheng-Jun Zha. 2024b. Escnet: Entity-enhanced and stance checking network for multi-modal fact-checking. In *Proceedings of the ACM on Web Conference 2024*, pages 2429–2440.

Fanrui Zhang, Jiawei Liu, Qiang Zhang, Esther Sun, Jingyi Xie, and Zheng-Jun Zha. 2023b. Ecenet: explainable and context-enhanced network for mutimodal fact verification. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 1231–1240.

Zhiwei Zhang, Jiyi Li, Fumiyo Fukumoto, and Yanming Ye. 2021. Abstract, rationale, stance: A joint model for scientific claim verification. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3580–3586.

Chen Zhao, Chenyan Xiong, Corby Rosset, Xia Song, Paul Bennett, and Saurabh Tiwary. 2020. Transformer-xh: Multi-evidence reasoning with extra hop attention. In *International Conference on Learning Representations*.

Yilun Zhao, Yitao Long, Tintin Jiang, Chengye Wang, Weiyuan Chen, Hongjun Liu, Xiangru Tang, Yiming Zhang, Chen Zhao, and Arman Cohan. 2024. Findver: Explainable claim verification over long and hybrid-content financial documents. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 14739–14752.

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

Wanjun Zhong, Jingjing Xu, Duyu Tang, Zenan Xu, Nan Duan, Ming Zhou, Jiahai Wang, and Jian Yin. 2020. Reasoning over semantic-level graph for fact checking. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6170–6180.

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.

Mingyang Zhou, Yi Fung, Long Chen, Christopher Thomas, Heng Ji, and Shih-Fu Chang. 2023. Enhanced chart understanding via visual language pretraining on plot table pairs. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1314–1326.

Anni Zou, Zhuosheng Zhang, and Hai Zhao. 2023. Decker: Double check with heterogeneous knowledge for commonsense fact verification. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 11891–11904.

A Mathematical Notations

Here we provide a summary of main mathematical notations used in the main paper in Table 8.

B Pseudo-Code of Training Process

We summarize the training process at Algorithms 1–2. We first optimize multi-modal evidence retriever by Algorithm 1. After convergence, we fix the parameters of the retriever and use it to produce retrieved evidence for claims. For claim verification and explanation generation in Algorithm 2, we jointly optimize both objectives using retrieved evidence in Algorithm 1.

Algorithm 1 Multi-Modal Evidence Retrieval

Input: A claim verification dataset $\mathcal{D}$ with claims $\mathcal{C}$, evidence texts $\mathcal{T}$, and images $\mathcal{I}$.

Output: Retrieved evidence for test claims.

1: Initialize model with pre-trained parameters in scientific domain or general domain.
2: **while** not converged **do**
3: Construct two-layer multi-modal graph for each claim with text and images and each evidence with text and images.
4: **for** evidence text t and images $i \in \mathcal{I}(t)$ **do**
5: Feature initialization by $\mathbf{H}_t^{(l=1)} = \text{TRM}(\mathbf{H}_t^{(l=0)})$ and $\mathbf{Z}_i^{(l=1)} = \text{ViT}(\mathbf{Z}_i^{(l=0)})$ and obtain token embeddings $\mathbf{H}_e^{(l=1)}$ for evidence sentence e.
6: **for** $l = 1, 2, ..., L-1$ **do**
// Image-to-text reasoning
7: Image-to-text reasoning by Eq. 3.
8: Transformer step with multi-modal multi-head attention Eq. 4
// Text-to-image reasoning
9: Text-to-image reasoning by Eq. 5.
10: ViT step with multi-modal multi-head attention Eq. 4.
11: **end for**
12: **end for**
13: Evidence embedding $\mathbf{h}_t = \mathbf{h}_{t,\text{CLS}}^{(L)}$.
14: Repeat Lines 4–12 and obtain claim embedding $\mathbf{h}_c = \mathbf{h}_{c,\text{CLS}}^{(L)}$.
15: Minimize loss function $\mathcal{L}_{\text{Ret}}$ in Eq. 7 with Adam optimizer.
16: **end while**

Algorithm 2 Multi-Modal Claim Verification and Explanation Generation

Input: A claim verification dataset $\mathcal{D}$ with claims $\mathcal{C}$, evidence texts $\mathcal{T}$, and images $\mathcal{I}$.

Output: Predicted veracity labels and generated explanations for test claims.

1: Initialize model with pre-trained parameters in scientific domain or general domain.
2: Use trained retriever in Algorithm 1 to output retrieved evidence for claims.
3: **while** not converged **do**
4: Obtain claim and evidence embedding matrices by multi-modal graph encoder Eq. 6.
5: **for** each claim c **do**
// Token-level fusion
6: Obtain unified claim and evidence embedding matrices $\mathbf{U}_c$ and $\mathbf{U}_{t_k}$ by Eqs. 8–10.
7: Claim-evidence interaction and obtain $\mathbf{U}_{c,t_k}$ by Eq. 11.
8: $\mathbf{U}_c = \text{mean}\big(\{\mathbf{U}_{t_k,c} | t_k \in \mathcal{T}(c)\}\big)$ and obtain claim embedding $\mathbf{c} = \mathbf{u}_{c,\text{CLS}}$.
// Evidence-level fusion
9: Obtain evidence embedding $\mathbf{t}$ by Eqs. 12–13.
10: $\hat{\mathbf{y}} = \text{softmax}(f_{\text{MLP}}([\mathbf{c}||\mathbf{t}]))$.
// Explanation generation
11: Obtain concatenated $\widehat{\mathbf{E}}_{c,t_k}$ by Eq. 15.
12: $\bar{\mathbf{E}}_c = \text{mean}(\tilde{\mathbf{E}}_{c,t_1}, ..., \tilde{\mathbf{E}}_{c,t_K})$.
13: $\hat{e} = f_{\text{LMDec}}(\bar{\mathbf{E}}_{c,t})$.
// Consistency regularization
14: For each word e_j, obtain logit $\mathscr{L}(e_j)$.
15: Mean pooling by Eq. 17.
16: $\hat{\mathbf{y}}_e = \text{softmax}(f_{\text{MLP}}(\mathscr{L}))$.
17: **end for**
18: Minimize overall loss function Eq. 19.
19: **end while**

Table 8: Summary of mathematical notations.

Notation	Description
$\mathcal{D}$	a claim verification dataset
$\mathcal{C}$	a set of $N = \lvert\mathcal{C}\rvert$ claims
$\mathcal{T}$	a corpus of $T = \lvert\mathcal{T}\rvert$ evidence texts
$\mathcal{I}$	a set of $I = \lvert\mathcal{I}\rvert$ images
$\mathcal{E}$	a set of $N = \lvert\mathcal{C}\rvert$ explanations for N corresponding claims
$\mathcal{Y}$	a set of labels
$\mathbf{h}_{t,\text{CLS}}^{(l)}$	evidence text t's [CLS] token embedding at the l-th step
$\mathbf{z}_{i,\text{CLS}}^{(l)}$	evidence image i's [CLS] token embedding at the l-th step
$\mathbf{u}_{c,\text{CLS}}$	unified multi-modal claim emebdding after token-level fusion, $\mathbf{c} = \mathbf{u}_{c,\text{CLS}}$
$\mathbf{t}$	unified multi-modal evidence embedding after token-level fusion
$\hat{\mathbf{y}}$	predicted label probability distribution from the verification module
$\mathbf{E}_c$	averaged multi-modal embedding matrix used to input to language model decoder for explanation generation
$\mathscr{L}(e_j\lvert\cdot)$	V-dimensional logits from language model decoder when generating token e_j
$\hat{\mathbf{y}}_e$	predicted label probability distribution from the explanation module

C Complexity Analysis

Here we analyze parameter count and complexity, and make a short comment on running time.

Computational Complexity. For evidence retrieval, we have $\mathcal{O}(L((W+2)d^2 + (W+2)^2 d + d^2))$ where W is the context length for positional encodings. For claim verification, we have $\mathcal{O}(W^2|\mathcal{I}(t)|Kd)$ where $|\mathcal{I}(t)|$ is the number of images each evidence text or claim text has, and K is the number of retrieved evidence. For explanation generation, we have $\mathcal{O}(W^2 dLK)$. Overall, the complexity is quadratic, which is a standard complexity for many language models (Vaswani et al., 2017; Devlin et al., 2019; Zhao et al., 2020; Raffel et al., 2020; Radford et al., 2021).

Parameter Count. For evidence retrieval, we have $(12Ld^2 + Vd + Wd) + (12Ld^2 + P^2C + Wd) + (d^2 + 2d)$ parameters. L is number of language modeling steps, d is the dimension of hidden embeddings, V is the size of language model vocabulary, W is the context length for positional encodings, P is the size of each patch in ViT, and C is the number of input channels. For claim verification, we have $8d^2$ parameters. For explanation generation, we have $28Ld^2 + Vd + Wd$ parameters.

Short Comment on Running Time. The main focus of our paper is on model effectiveness, not running efficiency. But for completeness, we still briefly report running time. On the largest dataset MR2, our model takes around 1.5 hours to converge. Experiments were conducted on 4 NVIDIA A100 GPUs. One possible method to improve running efficiency is to replace language model with LongFormer (Beltagy et al., 2020), which is efficient and supports long texts. Another method is to use distributed and parallel training on more and larger GPUs. We leave the research of optimizing running efficiency as a future work.

D Additional Discussion on Model Architecture

In this section we provide more explanations of the motivation behind model architecture to make the paper clearer and more self-contained.

Motivation of the GNN Module. For GNN, it supports variable number of neighbors for each node, and the main function of GNN is to aggregate neighboring nodes into a unified embedding. In our scenario, a text is associated with variable number of images. Although there are 5 or fewer images for each text, GNN can effectively aggregate images into a unified image embedding for language and vision modeling. Our design is similar to GEAR (Zhou et al., 2019), which also uses GNN to aggregate 5 or fewer evidence texts for fact-checking. Both GEAR and our experiments verify that GNN is an effective module for small-sized evidence graph. Our evidence graph is a multi-modal heterogeneous graph with both texts and images, and we use different matrices to project text embeddings and image embeddings to the same embedding space for GNN aggregation, please see Eq. 1 where we use $\mathbf{W}_{\text{txt}}$ and $\mathbf{W}_{\text{img}}$ for text and image, respectively.

Explanation on the Size of Multi-Modal Evidence Graph. Both our work and GEAR process an evidence graph with 5 or fewer nodes, and both papers show the effectiveness of applying GNN to process such an evidence graph. The additional

ablation study provided above also show that GNN aggregator is more effective than mean and max pooling aggregator. We would like to clarify that the purpose of GNN does not lie in how many neighbors a node has, but instead, the purpose lies in neighboring node aggregation with variable number of neighbors. The widely used Cora citation graph (Veličković et al., 2018) in graph community has 4 neighbors for each node on average, and GNNs are effective at learning node embedding on such a graph. Furthermore, as shown in GraphSAGE (Hamilton et al., 2017), the neighbor sampling strategy can sample as few as 2 neighbors for effective neighbor aggregation, and our work also uses the same neighbor sampling strategy to aggregate images of a text.

Motivation of Toekn-level and Evidence-level Fusion. For token-level and evidence-level fusion, the multi-modal graph encoder outputs embedding matrices for claims and evidence separately, and there is no interaction or reasoning between claim and evidence. We aim to use token- and evidence-level fusion to allow multi-modal claim and evidence to interact and reason with each other for claim verification.

Motivation of Multi-Modal Fusion-in-Decoder. For multi-modal fusion-in-decoder, we aim to aggregate multiple evidence texts with their images for explanation generation. Multi-modal fusion-in-decoder can effectively achieve this goal by concatenating image embedding with its corresponding evidence text embedding matrix, and by taking mean pooling for multiple pieces of evidence. The motivation of concatenation with image embedding is that we can incorporate image information into explanation generation. The motivation of taking mean pooling is that we effectively use the information of multiple evidence for explanation generation. Concatenating multiple evidence but not taking mean pooling is also possible, but it results in extremely long sequence, which is inefficient. Thus we choose mean pooling.

Motivation of Consistency Regularizer. For consistency regularizer, the motivation is similar to the design in MochegModel. We aim to make the generated explanation consistently reflect the predicted veracity, so that the explanation can well justify the reasoning process of how the model predicts the veracity.

E Additional Dataset Details

Here we provide dataset creation details of our introduced AIChartClaim, as well as data preprocessing details of other three publicly available datasets.

E.1 Details of AIChartClaim

Data source. For each paper, our annotators record publication year, venue, title, and caption of the chart. If the chart is not clearly readable, annotators search for other papers with appropriate charts. Some claims in the paper usually have a prefix, e.g., "Figure 1 shows that ...". To make the claims self-contained, our annotators remove this prefix and retain the main scientific findings in the sentence. Since the whole dataset is created by three different annotators, we have the fourth annotator to double check the whole dataset to make the creation consistent across all the claims and evidence.

Data augmentation. To create negations for the claims in the paper, we follow Wadden et al. (2020) and ask annotators to write negation for the 300 claims, taking precautions not to bias the negation by using obvious keywords, like "not". We mainly focus on the negation of semantic meaning of the claims. The "claim-only" version of our model, which removes both evidence texts and images for verification, achieves around 53%, which is quite close to 50%, suggesting that the negation process does not introduce severe artifacts.

To further augment the dataset by prompting GPT-4o for generation, for each chart with caption, we use below prompt to generate two more claims, one supported and one refuted by the chart:

Prompt for Claim Augmentation

> Please provide two more claims based on the chart and its caption that require multi-step reasoning, with one supported by the information in the chart and the other refuted by the information in the chart. The new claims should be different from previous claims. Meanwhile, please provide concise explanations in less than 100 words.

After generation, our annotators manually check and analyze the generated claims and explanations to make sure they are indeed correct, high-quality, and consistent with the charts. If there is any unclear or erroneous description, we either prompt GPT-4o to generate more claims with explanations, or manually correct the generated texts.

Explanation. We further generate explanations for those natural claims (i.e., claims obtained from

papers and their negations). For each chart with caption, we use below prompt for generation:

Prompt for Explanation Generation

The caption of this chart is "[CAPTION]". There is a claim based on the chart "[CLAIM TEXT]". Concisely explain why this claim is [SUPPORTED / REFUTED] by the chart in less than 100 words.

Here we replace [CAPTION] with the caption of the input chart image, and [CLAIM TEXT] with the claim text. We select either "SUPPORT" or "REFUTE" based on the label of the claim. Again, our annotators double check and analyze the generated explanation, and correct or regenerate it if there is any unclear or wrong description.

We split the created dataset into training:development:test by 70:10:20. For each publication venue, we randomly select 70% claims for training, 10% for validation, and 20% for testing. The split is balanced, and there is no overlap between training charts and test charts.

Human Evaluation. To further evaluate the quality of out dataset, we follow SciFact (Wadden et al., 2020) and randomly select 100 claim-chart pairs for re-annotation for three independent times. For each time, we randomly select different 100 claim-chart pairs and ask a different PhD student specializing in AI for re-annotation. These three PhD students were never involved in dataset creation before. The label agreement is 0.72 Cohen's κ (Wadden et al., 2020), which is a high agreement and is similar to the result in SciFact.

Motivation of Proposing AIChartClaim Dataset. We are motivated to create AIChartClaim dataset for three main reasons. *First*, most existing multi-modal fact-checking datasets are in the general domain. To our knowledge, our AIChartClaim dataset is the first scientific multi-modal fact-checking dataset in AI domain. Thus our dataset complements the research community. *Second*, understanding increases and decreases in quantities in scientific charts with AI domain-specific language is a crucial scientific reasoning ability for language models, thus it is necessary to have this scientific dataset. *Third*, we also make a comparison to other related datasets in Table 1. Existing scientific fact-checking datasets (SciFact, BearFact, and Check-COVID) are based on texts only, with no images. Existing multi-modal fact-checking chart datasets (ChartCheck and ChartFC) are in the general domain, and their content does not discuss scientific or more specifically, AI concepts. Other multi-modal fact-checking datasets (Mocheg, MR2, NewsCLIPpings, and FACTIFY) are general-domain datasets. They are neither chart nor scientific datasets. Such comparison shows that it is necessary to have our dataset to complement the research community.

E.2 Details of Other Datasets

ChartCheck[1] is a chart dataset in general domain with chart images and textual captions as multi-modal evidence. It also has explanations. Since the chart images in the dataset are provided by URL links, we accordingly download the charts using the links and remove charts and associated claims if the links are unavailable. In total, we have 1,615 charts with captions and 10,038 claims.

Mocheg[2] is a multi-modal dataset in general domain with explanations. We follow the instructions in GitHub and download the dataset. However, in their original paper (Yao et al., 2023), authors respectively use three different preprocessing methods for evidence retrieval, claim verification, and explanation generation, resulting in three different variations or subsets of the original dataset. In our model, we aim to keep consistent across different tasks, thus we keep all the claims, evidence, and images, and consistently use the same set of data for all three tasks. This is why our results, especially explanation generation, are slightly different from the results in the original paper, which uses only half of the claims for explanation generation. Finally, since some evidence texts do not have associated images, we use pre-trianed CLIP (Radford et al., 2021) to align texts and images. For each evidence text, we select top-3 most similar images and associate them with the text.

MR2[3] is another multi-modal dataset in general domain, but with no explanations. Both its claims and evidence texts have associated images. The downloaded dataset using the link in GitHub is slightly different from the one reported in the original paper in terms of dataset size (Hu et al., 2023). We tried very hard but still cannot obtain the same dataset. Thus, the results in our paper also have some deviations from the results in their original paper. The images of some evidence texts are provided in the dataset, while the images of other evidence texts are provided by a URL link.

[1] https://github.com/mubasharaak/ChartCheck
[2] https://github.com/VT-NLP/Mocheg
[3] https://github.com/THU-BPM/MR2

Table 9: Result of evidence retrieval with *Precision* score (%).

Model	AIChartClaim			ChartCheck			Mocheg			MR2		
	Prec@1	Prec@5	Prec@7	Prec@1	Prec@5	Prec@7	Prec@1	Prec@5	Prec@7	Prec@1	Prec@5	Prec@7
BM25	43.3±0.0	10.8±0.0	8.3±0.0	35.9±0.0	9.4±0.0	6.9±0.0	39.6±0.0	14.8±0.0	11.4±0.0	22.9±0.0	13.5±0.0	10.9±0.0
RAV	50.6±1.1	13.5±0.1	9.9±0.1	53.6±0.2	13.8±0.1	10.3±0.1	49.7±0.6	20.8±0.3	16.1±0.3	29.6±0.3	17.4±0.3	14.2±0.2
JustiLM	53.6±2.7	14.3±0.2	10.7±0.1	51.4±0.7	13.6±0.0	10.1±0.0	50.3±0.7	21.1±0.2	16.4±0.2	27.0±0.7	16.4±0.2	13.5±0.1
MochegModel	58.8±0.0	14.9±0.0	11.0±0.0	51.6±0.1	13.3±0.0	9.9±0.0	41.8±0.0	11.6±0.0	8.8±0.0	29.5±0.1	19.1±0.0	15.8±0.0
TransXH+ViT	53.3±3.5	14.2±0.5	10.6±0.2	53.2±0.2	13.8±0.2	10.3±0.1	51.1±0.7	21.1±0.2	16.4±0.2	35.5±1.2	22.9±1.0	18.3±0.9
MEVER w/o images	63.4±1.1	15.3±0.2	11.2±0.0	51.8±0.8	13.6±0.2	10.1±0.1	49.2±3.9	20.4±1.5	15.9±1.1	28.2±0.2	17.2±0.2	14.2±0.0
MEVER (ours)	**65.7±0.6**	**15.4±0.0**	**12.0±1.2**	**56.0±0.4**	**14.4±0.1**	**10.7±0.1**	**53.1±0.8**	**22.0±0.2**	**17.0±0.2**	**37.6±0.6**	**24.1±0.6**	**19.3±0.1**

Table 10: Result of evidence retrieval with *Recall* score (%).

Model	AIChartClaim			ChartCheck			Mocheg			MR2		
	Rec@1	Rec@5	Rec@7	Rec@1	Rec@5	Rec@7	Rec@1	Rec@5	Rec@7	Rec@1	Rec@5	Rec@7
BM25	43.3±0.0	54.2±0.0	57.9±0.0	35.9±0.0	46.9±0.0	48.5±0.0	17.9±0.0	30.8±0.0	32.9±0.0	3.8±0.0	10.6±0.0	11.9±0.0
RAV	50.6±1.1	67.4±0.6	69.4±1.0	53.6±0.2	69.2±0.5	72.0±0.3	22.8±0.4	43.0±0.7	46.4±0.8	5.0±0.1	13.8±0.3	15.7±0.4
JustiLM	53.6±2.7	71.5±0.9	74.6±0.4	51.4±0.7	67.8±0.1	70.7±0.2	23.4±0.3	43.8±0.5	47.0±0.5	4.5±0.1	13.1±0.1	14.9±0.1
MochegModel	58.8±0.0	74.6±0.0	77.1±0.0	51.6±0.1	66.7±0.0	69.1±0.1	22.2±0.0	30.2±0.0	31.6±0.0	4.9±0.0	14.9±0.0	16.9±0.0
TransXH+ViT	53.3±3.5	71.0±2.7	74.0±1.5	53.2±0.2	69.0±0.8	71.8±0.4	23.4±0.4	30.2±0.0	31.6±0.0	5.6±0.2	16.5±0.6	18.4±0.8
MEVER w/o images	63.4±1.1	76.5±0.9	78.2±0.2	51.8±0.8	67.8±0.9	70.5±0.9	22.6±1.7	42.3±2.1	45.6±3.2	4.8±0.0	13.7±0.0	15.7±0.1
MEVER (ours)	**65.7±0.6**	**77.2±0.2**	**79.0±0.5**	**56.0±0.4**	**72.2±0.5**	**75.0±0.5**	**24.4±0.4**	**45.2±0.4**	**48.7±0.4**	**5.9±0.1**	**17.3±0.3**	**19.2±0.4**

We accordingly download the images based on the link, and remove evidence text whose image link is unavailable. In total, we have 13,785 claims, 91,347 evidence texts, and 105,132 images.

F Experiment Environment

All the experiments were conducted on Linux server with 4 NVIDIA A100-SXM4-80GB GPUs. Its operating system is 20.04.5 LTS (Focal Fossa). We implemented our proposed model MEVER using Python 3.9 as programming language and PyTorch 2.4.1 as deep learning library. Other frameworks include numpy 1.24.1, sklearn 1.3.2, and transformers 4.46.0.

G Additional Experiment Results

Here we provide additional experiment results in terms of multi-modal evidence retrieval, claim verification, and explanation generation.

G.1 Evidence Retrieval

In the main paper, we report results of MAP, Precision@κ, and Recall@κ ($\kappa = 3$) for evidence retrieval. Here we provide additional results of precision and recall scores when varying κ in $\{1, 5, 7\}$ in Tables 9–10. Similarly, multi-modal baselines tend to outperform text-only baselines, verifying that images indeed bring useful information to improve the retrieval performance. Our model further outperforms multi-modal baselines, since we design a nested architecture to well integrate graph reasoning and vision-language modeling, thereby improving the result. The ablated version of our model, which removes images, deteriorates the retrieval performance, which further demonstrates that our model effectively incorporates images for multi-modal retrieval.

G.2 Claim Verification

In the main paper we report claim verification results with Macro F1 score. Here we provide additional results with Micro F1 score in Table **??**. Overall, multi-modal baseline models tend to outperform text-only baselines, since images provide auxiliary information to boost the verification accuracy. Our model further produces higher accuracy, since we model both token-level and evidence-level fusion to well reason between multi-modal claims and evidence, thereby achieving a more accurate verification. Though GPT-4o is slightly better than our model on the general-domain Mocheg dataset, our model still produces better results on Macro F1 score and on other datasets.

G.3 Explanation Generation

We present ROUGE-L, METEOR, and BLEU-2 scores with retrieved setting in the paper for explanation generation. Here we further provide ROUGE-1, ROUGE-2, and BLEU-4 scores with retrieved setting in Table 11. Similarly, we show the results of explanation generation with gold setting in Tables 12–13. Specifically, Table 11 presents additional evaluation metrics, i.e., ROUGE-1, ROUGE-2, and BLEU-4, for explanation generation with retrieved setting. Tables 12–13 show all six metrics of explanation generation with gold setting. Overall, we observe

that multi-modal baselines generate explanations more accurately than text-only baselines, since images complement texts for high-quality generation. Our model also outperforms baselines in most cases, showcasing the advantage of our multi-modal Fusion-in-Decoder and consistency regularizer.

To comprehensively evaluate explanation generation, we further adopt VLM-as-a-Judge for evaluation. Following existing works (Lee et al., 2024), we input below prompt to LLaVA-v1.5-7B (Liu et al., 2024c,b):

Prompt for VLM-as-a-Judge

USER: <image> You are an expert evaluator for fact-checking explanations. Please evaluate the quality of the following explanation, based on the provided claim, evidence text, evidence image, and predicted label. Your evaluation should consider the following four factors:

1. Label-Explanation Consistency: Does the explanation appropriately justify the predicted label?
2. Relevance: Is the explanation relevant to the claim, evidence text, and evidence image?
3. Correctness: Is the explanation factually correct based on the evidence text and evidence image?
4. Clarity: Is the explanation clear and easy to understand?

For each factor, assign a score from 1 (poor) to 5 (excellent).

Claim: [CLAIM TEXT]
Evidence Text: [EVIDENCE TEXT]
Predicted Label: [PREDICTED LABEL]
Explanation: [GENERATED EXPLANATION]

Please output your answer in the following format:
Label-Explanation Consistency: [SCORE]
Relevance: [SCORE]
Correctness: [SCORE]
Clarity: [SCORE]
ASSISTANT:

Table 14 shows the results. Overall, our model outperforms baselines in most cases. Although GPT-4o performs slightly better than our model for Correctness and Clarity on Mocheg dataset, our model still significantly produces better explanation on other evaluation dimensions and datasets. UniChart and ChartGemma are specifically designed for chart data, thus we do not show their result on Mocheg dataset.

G.4 Further Discussion on Three Evaluation Tasks

We use three tasks with multiple metrics to evaluate the performance of our model. There is no baseline model that performs very well on all tasks and metrics, while our model consistently performs better than or at least comparably with the best baseline on all tasks and metrics.

Specifically, for claim verification in Table 4, ChartGemma and Transformer-XH++ perform well on ChartCheck dataset, but on Mocheg and MR2 datasets Transformer-XH++ significantly deteriorates the performance (we conduct paired t-test with $p = 0.05$), and ChartGemma is even not designed for non-chart datasets. Similarly, GPT-4o and KGAT produce good result on Mocheg dataset, but on other datasets they fall behind our model with statistical significance (we conduct paired t-test with $p = 0.05$). This indicates that our model at least does not hurt the performance of any individual task or metric, but can significantly outperform baselines on many other tasks and metrics.

Similarly, for explanation generation, ChartGemma and ECENet produce good explanation on ChartCheck, but the performance of ECENet on Mocheg dataset is statistically significantly lower than our model's (again, ChartGemma is even not proposed for non-chart dataset, Mocheg). DePlot+FlanT5 performs well on Mocheg dataset, but on other datasets our model outperforms it with statistical significance. These results again verify that our model not only achieves better or comparable results on baselines' respective advantageous datasets, but significantly improves them on many other datasets, tasks, and metrics.

G.5 Failure Case and Error Type

We further examine the generated explanations of both our model and baselines. Here we summarize a main error type. Our model is built upon SciBERT (Beltagy et al., 2019) and T5 (Raffel et al., 2020), which can only capture limited length of evidence texts and explanations. Thus, for extremely long evidence texts and explanations that exceed the maximum length of SciBERT and T5, the quality of explanation generation may be influenced. Similarly, we observe the same performance drop for baseline models, which also have difficulty capturing extremely long sequence. One potential solution is that we can replace our language model with LongFormer (Beltagy et al., 2020), which can capture extremely long texts and maintain a good trade-off between model effectiveness and running efficiency.

Table 11: Explanation generation with *ROUGE-1*, *ROUGE-2*, and *BLEU-4* scores (%) with *retrieved evidence setting*.

Model	AIChartClaim			ChartCheck			Mocheg		
	ROUGE-1	ROUGE-2	BLEU-4	ROUGE-1	ROUGE-2	BLEU-4	ROUGE-1	ROUGE-2	BLEU-4
JustiLM	26.8±0.9	11.5±0.6	5.6±0.5	41.1±2.2	22.7±1.1	14.4±1.1	25.1±0.3	11.3±0.4	8.2±0.6
MochegModel	41.5±1.2	20.5±1.3	9.4±1.1	47.1±0.4	26.6±0.4	15.0±0.3	26.0±0.1	11.0±0.5	**10.7±0.1**
ECENet	40.9±0.6	20.5±0.8	10.6±0.5	46.9±0.3	26.0±0.2	15.1±0.2	24.9±0.4	11.1±0.5	8.1±0.2
DePlot+FlanT5	41.6±1.3	20.6±1.3	9.5±1.2	47.3±0.2	26.4±0.1	15.2±0.1	27.4±0.8	11.8±0.3	9.9±0.4
GPT-4o	24.1±0.6	18.4±0.5	7.1±0.2	22.5±0.2	8.9±0.6	12.7±0.2	**30.3±1.7**	9.9±1.7	4.5±0.8
UniChart	41.3±0.7	20.7±0.9	10.7±0.7	47.7±0.3	26.5±0.3	15.7±0.1	N.A.	N.A.	N.A.
ChartGemma	41.9±0.2	21.3±0.3	11.0±0.1	48.0±0.0	26.7±0.2	15.1±0.2	N.A.	N.A.	N.A.
MEVER w/o images	41.9±0.2	21.6±0.4	11.7±0.1	47.5±0.0	26.5±0.4	15.8±0.2	28.0±0.2	13.7±0.1	9.1±0.1
MEVER (ours)	**42.7±0.3**	**22.0±0.3**	**11.8±0.4**	**48.7±0.1**	**27.2±0.1**	**16.8±0.2**	28.5±0.3	**14.3±0.2**	10.1±0.2

Table 12: Explanation generation with *ROUGE-L*, *METEOR*, and *BLEU-2* scores (%) with *gold evidence setting*.

Model	AIChartClaim			ChartCheck			Mocheg		
	ROUGE-L	METEOR	BLEU-2	ROUGE-L	METEOR	BLEU-2	ROUGE-L	METEOR	BLEU-2
JustiLM	25.6±0.6	19.9±0.8	14.8±0.8	34.3±0.8	30.6±0.9	23.9±0.7	23.3±0.3	22.2±0.6	17.6±1.4
MochegModel	32.0±1.1	23.8±1.3	17.6±1.3	38.4±0.8	33.1±0.8	25.0±0.9	23.2±0.1	**22.7±0.2**	20.1±0.3
ECENet	33.1±0.2	26.0±0.1	19.1±0.2	40.0±0.3	34.2±0.2	25.7±0.2	22.4±0.2	21.3±0.4	14.6±0.2
DePlot+FlanT5	32.2±0.9	24.0±1.1	17.6±1.3	38.7±0.8	33.1±0.7	24.7±0.7	23.2±0.1	23.1±0.1	**20.6±0.2**
GPT-4o	18.7±0.4	23.4±0.3	16.1±0.5	18.4±0.7	28.5±0.7	16.9±2.3	17.8±0.7	18.8±0.8	11.7±0.4
UniChart	32.5±0.9	25.6±0.5	19.1±0.4	40.1±0.3	34.5±0.5	26.4±0.3	N.A.	N.A.	N.A.
ChartGemma	33.7±0.2	26.7±0.1	19.7±0.3	40.0±0.2	34.7±0.1	25.7±0.2	N.A.	N.A.	N.A.
MEVER w/o images	33.5±0.1	26.2±0.1	20.6±0.1	39.8±0.2	35.4±0.4	26.6±0.4	23.7±1.1	21.8±0.2	15.0±0.2
MEVER (ours)	**34.4±0.1**	**27.5±0.2**	**21.2±0.3**	**40.8±0.1**	**35.9±0.2**	**27.1±0.3**	**24.5±0.9**	**22.2±0.9**	16.6±0.2

Table 13: Explanation generation with *ROUGE-1*, *ROUGE-2*, and *BLEU-4* scores (%) with *gold evidence setting*.

Model	AIChartClaim			ChartCheck			Mocheg		
	ROUGE-1	ROUGE-2	BLEU-4	ROUGE-1	ROUGE-2	BLEU-4	ROUGE-1	ROUGE-2	BLEU-4
JustiLM	31.4±0.9	14.0±0.6	7.0±0.4	40.0±2.5	22.5±0.2	14.4±0.5	30.7±0.8	15.8±0.2	12.9±1.1
MochegModel	39.5±1.4	19.3±1.1	8.9±1.0	45.3±1.0	25.4±0.5	15.6±0.6	31.4±0.5	15.4±0.2	14.4±0.2
ECENet	41.5±0.2	20.4±0.1	10.2±0.1	47.4±0.3	26.4±0.4	15.2±0.1	29.2±0.3	15.2±0.1	12.4±0.1
DePlot+FlanT5	39.7±1.2	19.4±1.1	8.9±1.0	45.4±0.8	25.6±0.6	15.4±0.4	**32.1±0.2**	15.5±0.1	**14.7±0.1**
GPT-4o	25.5±1.0	18.1±0.8	6.8±0.2	23.0±1.2	9.6±0.0	13.2±0.1	31.2±2.5	10.1±0.5	4.7±0.1
UniChart	41.6±0.3	20.9±0.1	11.0±0.1	47.6±0.2	26.5±0.1	15.6±0.1	N.A.	N.A.	N.A.
ChartGemma	42.2±0.1	21.3±0.1	10.4±0.1	48.0±0.1	26.6±0.1	16.6±0.0	N.A.	N.A.	N.A.
MEVER w/o images	41.8±0.2	20.8±0.2	10.6±0.1	47.6±0.1	26.5±0.4	16.7±0.0	29.6±1.8	15.8±0.5	11.1±1.4
MEVER (ours)	**42.9±0.1**	**21.9±0.1**	**11.6±0.2**	**48.9±0.0**	**27.4±0.2**	**16.8±0.1**	**31.4±1.2**	**16.4±0.7**	12.4±0.9

Table 14: Explanation generation with *VLM-as-a-Judge*.

Model	AIChartClaim				ChartCheck				Mocheg			
	Consistency	Relevance	Correctness	Clarity	Consistency	Relevance	Correctness	Clarity	Consistency	Relevance	Correctness	Clarity
JustiLM	4.38±0.05	4.45±0.04	4.18±0.02	4.34±0.03	4.40±0.02	4.49±0.04	4.29±0.04	4.43±0.02	4.01±0.02	4.13±0.02	3.61±0.01	3.86±0.02
MochegModel	4.41±0.04	4.51±0.03	4.29±0.04	4.44±0.03	4.45±0.03	4.55±0.04	4.32±0.02	4.45±0.04	4.09±0.01	4.24±0.01	3.64±0.04	3.91±0.01
ECENet	4.42±0.02	4.52±0.02	4.28±0.03	4.41±0.02	4.44±0.05	4.54±0.04	4.30±0.03	4.44±0.04	4.09±0.01	4.23±0.01	3.64±0.02	3.90±0.01
DePlot+FlanT5	4.44±0.04	4.52±0.04	4.28±0.02	4.42±0.02	4.43±0.01	4.53±0.01	4.29±0.04	4.42±0.02	4.09±0.01	4.23±0.01	3.63±0.03	3.89±0.01
GPT-4o	4.29±0.04	4.42±0.03	4.19±0.03	4.21±0.00	4.28±0.02	4.47±0.02	4.20±0.01	4.23±0.02	4.10±0.02	4.24±0.01	**3.68±0.03**	**4.00±0.01**
UniChart	4.56±0.00	4.65±0.01	4.47±0.01	4.57±0.01	4.39±0.01	4.50±0.01	4.24±0.00	4.39±0.01	N.A.	N.A.	N.A.	N.A.
ChartGemma	4.60±0.00	4.68±0.00	4.53±0.01	4.62±0.01	**4.48±0.02**	4.58±0.01	**4.35±0.01**	4.48±0.01	N.A.	N.A.	N.A.	N.A.
MEVER w/o images	4.61±0.01	4.68±0.02	4.54±0.02	4.62±0.01	4.41±0.02	4.52±0.01	4.25±0.02	4.40±0.03	4.12±0.01	4.19±0.01	3.47±0.02	3.78±0.02
MEVER (ours)	**4.64±0.03**	**4.72±0.02**	**4.60±0.02**	**4.67±0.03**	**4.48±0.01**	**4.59±0.01**	**4.35±0.01**	**4.51±0.01**	**4.14±0.01**	**4.25±0.01**	3.66±0.01	3.97±0.02