Towards Interpretable Multimodal Fact Verification: A Hierarchical Prompting Framework with Large Vision–Language Models

Anonymous ACL submission

Abstract

The rapid spread of multimodal misinformation on online platforms poses significant challenges to automated fact verification, as textual claims are often tightly coupled with potentially misleading visual content. Existing multimodal fact verification approaches primarily rely on supervised, small-scale models, which exhibit limited reasoning ability and poor generalization in real-world scenarios. Although large vision–language models (LVLMs) demonstrate strong cross-modal understanding, they are not inherently optimized for fine-grained verification tasks and often produce unstable judgments when directly prompted. We propose **H**ierarchical **P**rompting **I**nterpretable **M**ultimodal Fact Verification (HPIM), an interpretable multimodal fact verification framework built on a hierarchical prompting strategy. The proposed method guides a large vision-language model through a coarse-to-fine reasoning process. This is achieved by first prompting a macro-level analysis of claims and evidence, followed by a micro-level, explanation-oriented analysis that leverages structured factual elements. Subsequently, the method fuses textual, visual, and analytical representations to predict veracity. This final prediction is then fed back into the model, enabling it to generate explanations grounded in the evidence. Experiments on a public benchmark demonstrate strong verification performance and improved interpretability. Code is available at: https://anonymous.4open.science/r/HPIM-74D9.

1 Introduction

The rapid expansion of online media ecosystems has substantially accelerated the dissemination of information, while simultaneously facilitating the large-scale spread of multimodal misinformation that intertwines textual claims with visual content. Such content often presents claims in a persuasive yet deceptive manner, making it difficult for users to assess their authenticity. The unchecked propagation of multimodal misinformation can distort public understanding, intensify social tensions, and undermine trust in authoritative information sources, thereby highlighting the pressing need for robust automated multimodal fact verification systems.

Fact verification aims to assess the veracity of a claim by systematically examining its consistency with reliable evidences. In today's online environment, evidence is increasingly multimodal. Images are not just illustrations but integral components that can either corroborate or contradict the text. Existing approaches for multimodal fact verification largely rely on supervised, small-scale multimodal language models (Hu et al., 2022; Wang et al., 2024a; Mu et al., 2024). While these models demonstrate promising performance under controlled or in-distribution settings, they often exhibit limited effectiveness in real-world fact-checking scenarios. This limitation stems from restricted knowledge coverage, insufficient reasoning capabilities, and weak generalization to novel events, entities, or domains.

Recent advances in LVLMs have introduced new opportunities for multimodal fact verification. Owing to pretraining on large-scale multimodal corpora, LVLMs possess strong cross-modal alignment and reasoning abilities, enabling them to jointly interpret textual claims and visual evidence (Xu et al., 2024; Zhao et al., 2023). However, despite their general-purpose strengths, LVLMs are not inherently tailored for fine-grained verification tasks that require precise judgment of claim–evidence relationships (Nan et al., 2024). Directly prompting LVLMs to determine news authenticity often results in unstable predictions and limited interpretability, which constrains their practical adoption in fact-checking pipelines.

Motivated by the structured manner in which human fact-checkers progressively analyze

news—from high-level event plausibility to fine-grained evidence inspection—we propose HPIM, a practical and interpretable framework for multimodal fact verification. At the core of HPIM is a hierarchical prompting structure that explicitly decomposes the verification process into multiple reasoning stages and guides LVLM through coarse-to-fine analysis. The framework first applies a macro-level prompt to elicit a global assessment of the claim, focusing on essential event elements such as entities, events, locations, and temporal information. These extracted elements are then incorporated into micro-level prompts that direct the LVLM to examine detailed claim–image consistency, enabling targeted cross-modal reasoning grounded in specific factual components. The resulting hierarchical analyses, together with the original claim and visual evidence, are integrated into a lightweight multimodal verification module. This module employs an attention-based fusion mechanism to model fine-grained interactions across modalities and produce a final veracity prediction. To enhance transparency and trust, the predicted label is subsequently fed back into the LVLM through an explanation-oriented prompt, generating an evidence-grounded, human-readable rationale that explicitly reflects the hierarchical reasoning process.

Our contributions are summarized as follows:

- We propose HPIM, a multimodal fact verification framework that utilizes a hierarchical prompting structure to conduct a multi-level analysis, examining information from both macro and micro perspectives to improve verification accuracy.
- We employ structured news element extraction to isolate key entities from the text, enabling a fine-grained alignment of these entities with visual evidence for enhanced verification accuracy and evidence-based explainability.
- Experiments on public benchmark datasets demonstrate that HPIM achieves strong performance and provides more informative, human-readable rationales for multimodal fact verification.

2 Related Work

2.1 Multimodal Fact Verification

Existing approaches to multimodal fact verification are largely derived from supervised multimodal fake news detection frameworks, which focus on learning joint representations of textual claims and visual evidence to support veracity classification (Liu et al., 2023b; Chen et al., 2023). These methods typically rely on pretrained unimodal encoders—such as Vision Transformer (ViT) for images (Dosovitskiy, 2020) and RoBERTa for text (Liu et al., 2019)—to extract modality-specific features, which are subsequently combined through manually designed fusion mechanisms to produce a final verification decision (Zhou et al., 2020; Zhang et al., 2021). Recent studies have shifted their focus to modeling semantic consistency across modalities for finer-grained fact verification. Representative works, such as CAFÉ (Chen et al., 2022) and FND-CLIP (Zhou et al., 2023), enhance the precision of identifying image-text mismatches by quantifying semantic ambiguity or using element-level semantic alignment. Additionally, some approaches use Graph Neural Networks (GNNs) to integrate external knowledge (Hu et al., 2021; Wang et al., 2020).

Despite their performance gains, these methods struggle with modeling entity-level modality alignment and representing subject-object interactions, which hinders fine-grained reasoning. To address this, we propose an LVLM-based framework that uses structured news element extraction to align key textual entities with visual evidence. This approach enhances the model's capacity to identify subtle inconsistencies in news reports.

2.2 Large Vision-Language Models

Recent advancements in LVLM, such as LLaVA (Liu et al., 2023a) and Mini-GPT4 (Zhu et al., 2024), have showcased their potential in multimodal tasks (Chiang et al., 2023; Touvron et al., 2023; Wang et al., 2024b). A typical LVLM architecture consists of an image encoder, a projector, and a Large Language Model (LLM). The projector converts visual features into visual prompt embeddings, which are combined with text prompts and fed into the LLM. This architecture has spurred interest in applying the reasoning capabilities of LVLMs to fake news detection (Tahmasebi et al., 2024; Zheng et al., 2025).

However, standard LVLMs are often less accurate for this purpose than smaller, specialized models because they lack the detailed focus required for fact verification. To solve this, researchers have tried various strategies, such as adding forgery-specific knowledge (Liu et al., 2024) or breaking the task into smaller sub-tasks (Wan et al., 2024).

Despite progress, current LVLM-based methods for fact verification remain less accurate and robust than specialized models. They often overlook the power of prompts, failing to create fine-grained prompt embeddings that guide the LLM to spot critical cross-modal inconsistencies and entity-level details. This significantly limits their effectiveness. To address this, We propose a novel framework for creating fine-grained prompt embeddings that empower an LLM to identify critical cross-modal inconsistencies and entity-level details.

3 Proposed Method

In this section, we introduce HPIM, a multimodal fact verification framework centered on a hierarchical prompting structure that explicitly decomposes the verification process into multiple reasoning stages and guides LVLM through structured, coarse-to-fine analysis. As illustrated in Figure 1, HPIM operates through three sequential stages. First, it uses hierarchical prompts to guide a vision-language model through macro and micro analyses of claim-image consistency. Then an attention-based fusion mechanism is employed to model fine-grained interactions across modalities and produce a final veracity prediction. Finally, the predicted label is subsequently fed back into the LVLM through an explanation-oriented prompt, generating an evidence-grounded, human-readable rationale that explicitly reflects the hierarchical reasoning process.

3.1 LVLM Analysis Module

This module is designed to conduct a preliminary analysis of news reports and their associated evidence. First, the elemental components of a news claim are extracted. Subsequently, a manually constructed prompt for macro-level analysis is provided to the vision-language model, enabling the model to generate an initial analytical summary of the news content.Finally, CLIP (Radford et al., 2021), together with learnable prompts, is employed for a micro-level consistency examination of the multimodal information.

3.1.1 Extraction of News Elements

Extracting key news elements—such as event time, location, subjects, and objects—is a critical first step that directly enhances both detection accuracy and interpretability. Rather than performing a coarse comparison between raw text and an image, this structured extraction enables fine-grained, cross-modal verification; for instance, the model can explicitly check if a person mentioned in the text is visually present in the image. This process allows the model to move beyond simple narrative plausibility and assess the logical and factual consistency between textual claims and visual evidence. Ultimately, these extracted elements serve as the essential building blocks for both the subsequent attention fusion module to make a precise authenticity judgment and for the final explanation module to generate concrete, human-interpretable rationales.

Accordingly, this study employs the Stanford NLP toolkit (Manning et al., 2014) for Named Entity Recognition (NER) to automatically extract key information (such as time, location, subject, object, and causal structures) from news texts, thereby establishing a basis for analyses of logical consistency and truthfulness. For a given claim C, NER is applied to obtain a list of entities, expressed as:

$$\begin{aligned} C^L &= \mathrm{NER}(C) \\ &= [C_i^{\mathrm{type}}, C_i^{\mathrm{text}}] \mid i = 0, 1 \ldots \end{aligned} \tag{1}$$

where C^L denotes the entity list of C, $\mathrm{NER}(*)$ represents the named entity recognition operation. C_i^{type} and C_i^{text} denote the entity type and entity text of the i-th entity in C, respectively.

3.1.2 Macro-Level Analysis with a Guiding Prompt

The initial stage involves a macro-level analysis of the news content. To achieve this, we employ prompt engineering to construct a comprehensive input for the LVLM. This input is formatted as follows:

$$input = \mathrm{Format}(C, C^L, E_V, E_T, P) \tag{2}$$

Here, E_V and E_T are the associated visual and textual evidence, respectively. P is a manually designed prompt template that directs the LVLM to perform a broad, initial analysis of these core components. The $Format(*)$ function organizes these elements into a coherent input, and the LVLM processes this to generate a preliminary analytical summary, denoted as A. The discussion and details regarding the prompts are presented in the appendix A.1.

3.1.3 Micro-Level Encoding with Learnable Prompts

Following the macro-analysis, we proceed to a micro-level encoding of the multimodal informa-

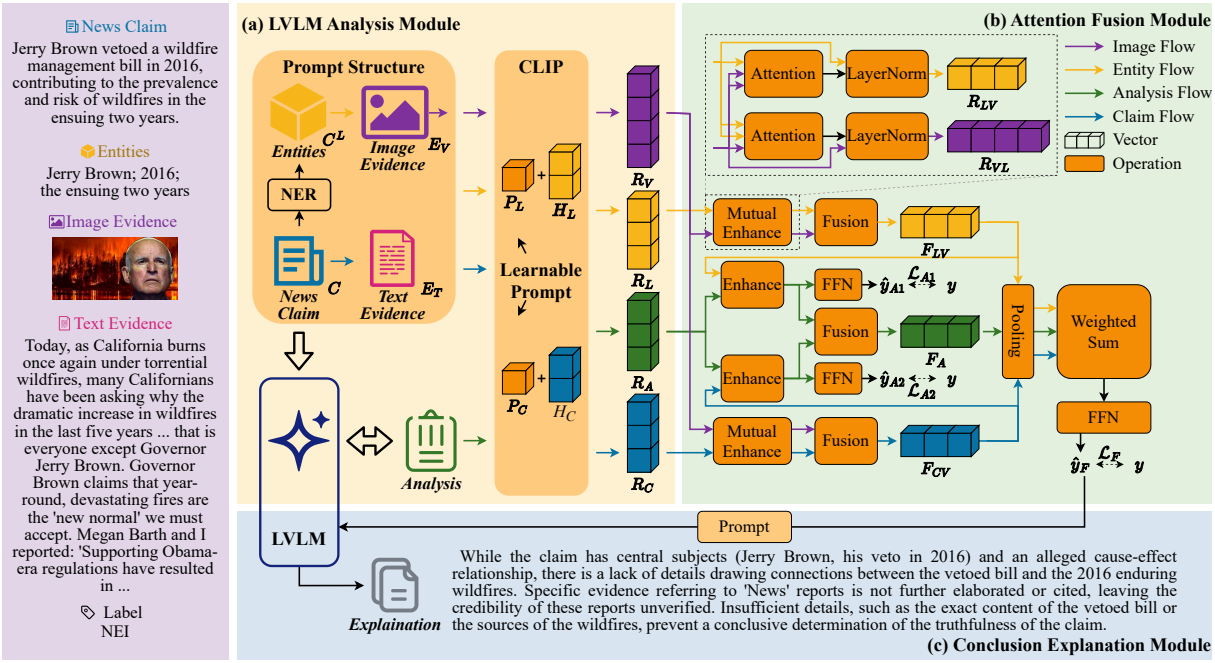


Figure 1: A schematic overview of HPIM. For each news claim and its corresponding evidence (both image and text), HPIM processes the input in three stages: (a) It extracts key news elements and analyzes them using LVLM to produce an analytical summary. (b) It encodes the textual and visual content with CLIP, and then applies an attention-based fusion mechanism to integrate multimodal features and generate a prediction of the news item's veracity. (c) It employs LVLM once more to provide a coherent and justified explanation for the veracity assessment.

tion using CLIP (Radford et al., 2021), which is known for its rich semantic embedding space. This stage integrates a second layer of prompting through learnable prompt tokens.

Claim and Entity Encoding: For the claim C and its key entities C^L, we first use the CLIP text encoder to extract their feature sequences (H_C and H_L). Inspired by Prompt Tuning, we introduce dedicated sequences of learnable prompt tokens, P_C and P_L. These learnable prompts are prepended to their respective feature sequences (e.g., $[P_C; H_C]$). The augmented sequences are then passed through CLIP's text projection layer to produce the final, fine-tuned representations, R_C and R_L.

Visual and text Encoding: In this stage, we generate distinct representations for the evidence and its analysis. The raw visual evidence E_V is encoded using CLIP's powerful image encoder. Meanwhile, the synthesized analytical passage A from the LVLM is encoded via the corresponding text encoder. Both are then projected into the final feature space to obtain their definitive representations, R_V and R_A.

3.2 Attention Fusion Module

This module is designed to provide a judgment of news veracity. First, CLIP, together with learnable prompts, is employed to encode the multimodal content. An attention-based fusion mechanism is then used to adaptively integrate the claim, image evidence, and entity features. In addition, the initial analytical passages generated by Section 3.1 are incorporated to obtain the final representation, which is subsequently fed into a classifier to generate predictions.

3.2.1 Multimodal Information Fusion

Comparing entities extracted from a textual claim with those depicted in accompanying images enables the identification of evidence-based cues for assessing claim veracity. For instance, one may verify whether the principal individuals described in the news text are consistent with those depicted in the image. Therefore, to model the relationship between news entities and evidential images, we employ multi-head cross-attention to integrate these multimodal sources of information.

Let R_L represent prompt-augmented entity representation, R_V the evidential image, H the number of heads in the multi-head cross-attention module, and d the model dimensionality. The enhanced representations for prompt-augmented entity is computed as follows:

$$\mathrm{En}^{L} = \mathrm{Attention}(R_L, R_V, R_V) \quad (3)$$
$$= \left(\Big\|_{h=1}^{H} \mathrm{softmax}\left(\frac{\mathrm{Q}^{L}{}_{h} \mathrm{K}^{V\top}_{h}}{\sqrt{d/H}} \right) \mathrm{V}^{V}{}_{h} \right) \mathrm{W}^{L}{}_{O} \Bigg)$$

where "$\|$" denotes concatenation operation. $\mathrm{Q}^{L}{}_{h}$, $\mathrm{K}^{V}{}_{h}$, and $\mathrm{V}^{V}{}_{h}$ are obtained from R_L and R_V through parameterized transformation matrices. Here, Q, K, and V denote the query, key, and value in the attention mechanism, respectively. While $\mathrm{W}^{L}{}_{O} \in \mathbb{R}^{d \times d}$ represents the final linear transformation layer of the model. Similarly, the representation En^{v} can be derived in the same manner. Subsequently, we apply residual connections followed by LayerNorm:

$$R_{LV} = \mathrm{LayerNorm}(R_L + \mathrm{En}^{L}) \quad (4)$$

$$R_{VL} = \mathrm{LayerNorm}(R_V + \mathrm{En}^{V}) \quad (5)$$

In this way, we obtain representations in which news entities and visual evidence mutually enhance each other.

Given the dimensional mismatch between R_{LV} and R_{VL}, we apply an adaptive average pooling operation to generate R'_{LV} and R'_{VL}, ensuring that their dimensions correspond to those of R_{VL} and R_{LV}, respectively, thereby facilitating subsequent feature fusion.

$$R'_{LV} = \mathrm{AdaptivePool}(R_{LV}) \quad (6)$$

$$R'_{VL} = \mathrm{AdaptivePool}(R_{VL}) \quad (7)$$

Thereafter, features of news entities and visual evidence are extracted via an attention mechanism and subsequently integrated through learnable weighting parameters:

$$F_{LV} = \mathrm{Fusion}(R_{LV}, R_{VL}) = w_1 f_1 + w_2 f_2 \quad (8)$$

$$f_1 = \mathrm{Attention}(R'_{LV}, R_{VL}, R_{VL}) \quad (9)$$

$$f_2 = \mathrm{Attention}(R'_{VL}, R_{LV}, R_{LV}) \quad (10)$$

where F_{LV} encompasses the semantic features of news entities and visual evidence, with w_1 and w_2 denoting trainable weights. Considering that R_L includes only news entities rather than complete news sentences—and thus lacks the causal logic encoded in full sentences—we apply the same procedure to R_C and R_V, ultimately obtaining F_{CV}.

3.2.2 Analytical Rationales Fusion

Since LVLMs inevitably exhibit hallucinations, the semantic information contained in R_A does not always contribute effectively to determining the veracity of news. F_{LV} captures the relational coherence among news entities, while F_{CV} emphasizes the logical consistency of news contexts. We integrate each of these components with R_A to more fully uncover the underlying rationales embedded within R_A. Finally, we can get F_A which represents the fused features that incorporate the analytical rationales.

$$F_A = \mathrm{LayerNorm}(R_A + \mathrm{Attention}(R_A, F_{LV}, F_{LV})) + \mathrm{LayerNorm}(R_A + \mathrm{Attention}(R_A, F_{CV}, F_{CV})) \quad (11)$$

3.2.3 Prediction

Building on the preceding outputs, we employ an adaptive-weighting strategy to integrate them and produce the final prediction. To reduce the dimensionality of the features, we first apply average pooling along the last dimension and subsequently fuse the resulting representations.

$$F'_A = \mathrm{AvgPool}(F_A) \quad (12)$$

$$F = \alpha(\beta F'_{LV} + (1-\beta) F'_{CV}) + (1-\alpha) F'_A \quad (13)$$

where F denotes the final representation used for prediction, with F'_{LV}, F'_{CV}, and F'_A representing the pooled features of F_{LV}, F_{CV}, and F_A, respectively. Here, α and β are trainable weights that balance the contributions of each component. The final prediction is obtained using a classifier constructed from a feedforward neural network with multiple hidden layers. The predicted outputs are then compared with the ground-truth labels using the cross-entropy loss function:

$$\hat{y}_F = \mathrm{FFN}(F) \quad (14)$$

$$\mathcal{L}_F = \mathrm{CrossEntropyLoss}(\hat{y}_F, y) \quad (15)$$

where $\hat{y}_F$ denotes the prediction results, y represents the ground-truth labels. In addition, to evaluate the effectiveness of LVLM in generating analytical passages, two independent feedforward neural networks are employed to make predictions based on the analytical rationale features:

$$\hat{y}_{A1} = \mathrm{FFN}(R_{A1}), \hat{y}_{A2} = \mathrm{FFN}(R_{A2}) \quad (16)$$

$$\mathcal{L}_{A1} = \mathrm{CrossEntropyLoss}(\hat{y}_{A1}, y) \quad (17)$$

$$\mathcal{L}_{A2} = \text{CrossEntropyLoss}(\hat{y}_{A2}, y) \quad (18)$$

the final loss function is obtained by computing a weighted sum of these individual loss terms:

$$\mathcal{L}_{final} = \mathfrak{a}_1 \mathcal{L}_F + \mathfrak{a}_2 \mathcal{L}_{A1} + \mathfrak{a}_2 \mathcal{L}_{A2} \quad (19)$$

where $\mathfrak{a}_1$, $\mathfrak{a}_2$ ($\mathfrak{a}_2 = 0.5 \times (1 - \mathfrak{a}_1)$) are hyperparameters that balance the contributions of each loss component.

3.3 Conclusion Explanation Module

After obtaining the prediction results from the attention fusion module, we feed the prediction $\hat{y}_F$ back into the LVLM to generate an explanatory passage. The input to the LVLM is constructed as follows:

$$input = \text{Format}(C, C^L, E_V, A, \hat{y}_F, P') \quad (20)$$

where P' denotes a manually designed prompt template that guides the LVLM to generate explanations, different from P in Section 3.1.

The discussion and details regarding the prompts are presented in the appendix A.2.

4 Experiments

4.1 Experimental Settings

Datasets. In our experiments, we employed the MOCHEG dataset (Yao et al., 2023) as the benchmark to comprehensively evaluate the performance of the proposed HPIM approach in the task of fact verification. This dataset comprises 15,601 labeled news claim samples (supported, refuted, NEI), accompanied by 33,880 corresponding textual evidence pieces and 12,112 visual evidence items. Its substantial scale and high degree of diversity align well with the practical requirements of fact verification.

Baseline. We selected the following methods as baseline models for evaluation: GEAR (Zhou et al., 2019), HESM (Subramanian and Lee, 2020), KGAT (Liu et al., 2020), Triple-Check (Du et al., 2022), Ino (Zhang et al., 2023), Logically (Verschuuren et al., 2023), MOCHEG (Yao et al., 2023), HGTMFC (Pang et al., 2025), EExpFND (Wang et al., 2025) and MSP (Chen et al., 2024). Among these, GEAR, HESM, and KGAT are unimodal approaches that rely solely on textual information, while the remaining methods are multimodal approaches that leverage both text and images. The evaluation metrics include Accuracy (Acc), Precision (Pre), Recall (Rec), and F1 score (F1), enabling a comprehensive assessment of model performance.

Implementation Details. During training, we adopted the Adam optimizer with a learning rate of 2×10^{-5}, a weight decay of 0.01, and a batch size of 16. In addition, the hyperparameters $\mathfrak{a}_1$ and l were set to 0.7 and 8, respectively. All implementations were developed using the PyTorch framework and executed on an NVIDIA RTX 4090 24GB GPU.

4.2 Performance of HPIM

We evaluate HPIM on the MOCHEG dataset and compare it against existing multimodal fact verification approaches. The main experimental results are summarized in Table 1. As indicated by the empirical findings, the proposed HPIM achieves substantially better performance than existing models on the MOCHEG dataset.

In terms of quantitative results, Our proposed HPIM achieves an accuracy of 59.94%, representing a 4.11% improvement over the previously best-performing model. Furthermore, the F1 scores for the refuted, supported, NEI categories increased by 5.86%, 0.16%, and 3.26%, respectively. These findings demonstrate that HPIM exhibits strong advancement and practical utility for multimodal news veracity recognition tasks.

4.3 Ablation Study

In this section, we discuss the ablation studies. To investigate the contribution of each component in HPIM, we remove the following modules individually: -w/o P_t: This variant eliminates the learnable prompt tokens, P_C and P_L, meaning that the claim and entity representations are directly fed into the model for multimodal fusion. -w/o R_A: This variant removes the analytical passage input together with all associated processing steps, effectively excluding the LVLM component to assess its role within HPIM. -w/o R_L: This variant discards the entity feature input and replaces it with a zero vector. -w/o R_C: This variant removes the claim feature input and substitutes it with a zero vector. -w/o R_V: This variant removes the visual evidence feature input and replaces it with a zero vector.

The experimental results presented in Table 2 clearly delineate the specific effects of different feature inputs and architectural components on overall model performance. Notably, when the raw news content is removed (i.e., the -w/o R_C variant), the

	Acc(%)	Refuted			Supported			NEI		
		Pre(%)	Rec(%)	F1(%)	Pre(%)	Rec(%)	F1(%)	Pre(%)	Rec(%)	F1(%)
GEAR	41.97	53.02	52.91	52.96	38.01	38.95	38.48	25.35	24.51	24.93
HESM	48.46	59.72	62.62	61.14	44.99	42.88	43.91	29.20	28.40	28.80
KGAT	48.55	59.69	63.11	61.35	45.24	42.88	44.03	28.89	27.82	28.34
Logically	48.51	58.63	61.65	60.10	46.91	42.61	44.66	29.94	30.93	30.43
Ino	48.59	59.07	61.94	60.47	46.25	41.96	44.00	30.47	31.71	31.08
Triple-Check	49.07	59.48	62.72	61.06	47.29	42.22	44.61	30.37	31.91	31.12
MOCHEG	52.06	61.50	65.44	63.41	53.19	47.97	50.45	30.78	31.32	31.05
HGTMFC	54.09	62.75	68.35	65.43	55.67	48.76	51.99	33.27	33.46	33.37
EExpFND	54.96	62.94	68.25	65.49	56.81	50.72	53.59	34.97	34.63	34.80
MSP	55.83	63.43	67.86	65.57	57.63	**53.33**	55.40	36.47	35.41	35.93
HPIM	**59.94**	**66.20**	**77.57**	**71.43**	**58.94**	52.55	**55.56**	**43.57**	**35.60**	**39.19**

Table 1: Performance comparison of HPIM with baseline methods on the MOCHEG dataset. The best results are highlighted in bold.

	Acc(%)	**Pre(%)**	**Rec(%)**	**F1(%)**
HPIM	**59.94**	**56.24**	**55.24**	**55.39**
-w/o P_t	56.34	53.10	52.90	52.71
-w/o R_A	58.94	55.94	52.84	52.69
-w/o R_L	56.47	53.81	50.69	50.14
-w/o R_C	53.23	50.40	49.85	49.70
-w/o R_V	58.51	54.17	52.54	52.41

Table 2: Ablation study results of HPIM on the MOCHEG dataset.

model exhibits the poorest performance across all evaluation metrics, falling significantly below the baseline. This finding underscores the essential role of original news text as a fundamental source of information. It further indicates that relying solely on analytical passages generated from image content by LVLM is insufficient for capturing event context or preserving semantic completeness, thereby constraining the model's ability to understand complex news narratives.

Meanwhile, both the -w/o P_t variant, which removes the learnable prompt tokens, and the -w/o R_L variant, which removes the entity extraction mechanism, show notable performance degradation. These results highlight the importance of learnable prompts in guiding the model to attend to critical information, as well as the indispensable role of entity recognition in producing structured semantic representations. Together, these two components constitute the core mechanism through which the HPIM approach effectively models multimodal news content.

In addition, the -w/o R_A and -w/o R_V variants also demonstrate varying degrees of performance decline, further corroborating the crucial role of LVLM in image understanding and semantic abstraction. This not only emphasizes the contribution of visual analytical information to the model's discriminative capability, but also provides empirical evidence that visual evidence must still participate in subsequent multimodal fusion after being processed by LVLM. Such participation facilitates deeper semantic alignment and complementarity between visual and textual information.

4.4 Evaluation on Explanations

To assess the quality of generated explanations, we conducted a blind human study with 10 volunteers who just rated the outputs based on specific criteria, as automatic metrics often fail to capture semantic logic (Chang et al., 2024). We randomly sampled 60 explanations for evaluation. For comparison, we selected three models that represent the current spectrum of capabilities in multimodal and language understanding, including LLaVA(Li et al., 2024), ChatGPT (Achiam et al., 2023), and Gemini (Comanici et al., 2025). To maintain fairness, we unified the prompt constraints across all models, explicitly specifying the response length to ensure consistency. We employed a 5-point Likert scale across four subjective dimensions, complemented by one objective metric to validate the capture of news elements:

- **Informativeness (I):** Evaluates whether the explanation provides rich background context and specific new information.
- **Faithfulness (F):** Assesses whether the explanation aligns with the ground truth and is free from hallucinations.
- **Completeness (C):** Measures whether the explanation covers all necessary aspects for news verification.

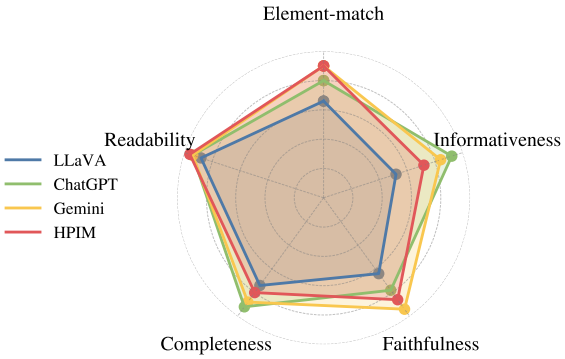


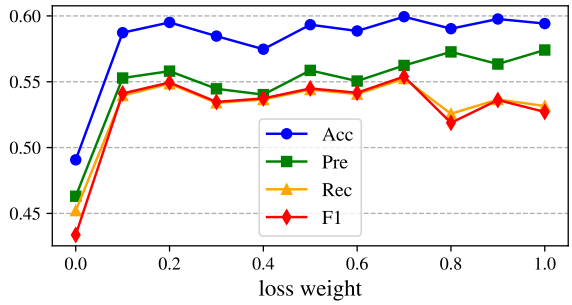
Figure 2: Performance comparison of generated explanations.

- **Readability (R):** Checks for grammatical correctness and sentence coherence.
- **Element-match (E):** An objective metric calculating the recall rate of named entities in the explanation against the source text.

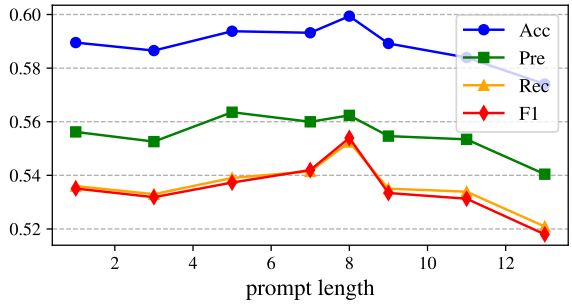
As shown in Figure 2, our model outperforms all baselines in element-match, validating its superior ability to ground explanations in specific news elements. While slightly trailing ChatGPT and Gemini in informativeness and completeness, our method achieves comparable performance in faithfulness and readability. Considering that ChatGPT and Gemini are massive, web-connected models with external knowledge access, our locally deployed model demonstrates exceptional efficiency, achieving high-quality, hallucination-free explanations under significant resource constraints.

4.5 Hyperparameter Sensitivity Analysis

In this section, we conduct a sensitivity analysis of HPIM regarding two key hyperparameters: the weighting factor α_1 used to construct the final loss function, and the prompt length l. As illustrated in Figure 3(a), with the exception of the extreme case where $\alpha_1 = 0$, different hyperparameter settings impose only marginal effects on HPIM. We attribute this phenomenon to the adaptive integration of entity relations, contextual logic, and multimodal representations, which renders the model less sensitive to variations in α_1. Figure 3(b) presents the sensitivity of HPIM to the prompt length l. We observe that the performance of HPIM initially improves as l increases, but subsequently degrades, achieving its peak when l is in the range of 7 to 8. This phenomenon likely stems from a trade-off: while increasing l provides expanded learnable context, an excessive length may inadvertently introduce noise.



(a) Performance vs. α_1



(b) Performance vs. l

Figure 3: Performance of HPIM under different hyperparameter settings on the MOCHEG dataset.

5 Conclusion

This paper proposes a novel multimodal fact verification framework that leverages LVLM through a hierarchical, coarse-to-fine prompting strategy to improve both accuracy and interpretability. The method first extracts structured news elements (e.g., entities, time, location, and causal relations) and guides an LVLM to perform macro-level semantic analysis, then applies CLIP-based micro-level encoding with learnable prompts and an attention-based fusion module to integrate textual, visual, and analytical features for veracity prediction. Finally, the predicted result is fed back into the LVLM to generate human-readable, evidence-grounded explanations. Experiments on the MOCHEG benchmark demonstrate that HPIM outperforms existing unimodal and multimodal baselines, while also producing more faithful and interpretable rationales, highlighting its practical value for robust multimodal news verification.

Limitations

One limitation of our work is the high time cost associated with LVLMs. Unlike traditional methods, generating detailed analyses and explanations for

large-scale datasets is computationally intensive and slow. However, we anticipate that rapid advancements in model efficiency will soon alleviate this bottleneck. Currently, we prioritize the depth and interpretability of the generated insights over processing speed.

Ethical Considerations

This research focuses exclusively on defensive strategies to assist content moderation. While we acknowledge the theoretical possibility of adversarial adaptation—whereby actors might attempt to bypass detection—we advocate for human-in-the-loop deployment to mitigate such risks. Regarding human evaluation, we enforced strict safety protocols to safeguard participant mental health against sensitive content. Finally, data processing was restricted to de-identified records, conducted in full compliance with platform regulations and privacy governance to ensure no personal information was compromised.

References

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

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, and Yidong Wang. 2024. A survey on evaluation of large language models. *ACM transactions on intelligent systems and technology*, 15(3):1–45. ISBN: 2157-6904 Publisher: ACM New York, NY.

Ting-Chih Chen, Chia-Wei Tang, and Chris Thomas. 2024. Metasumperceiver: Multimodal multi-document evidence summarization for fact-checking. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8742–8757.

Yixuan Chen, Dongsheng Li, Peng Zhang, Jie Sui, Qin Lv, Lu Tun, and Li Shang. 2022. Cross-modal ambiguity learning for multimodal fake news detection. In *Proceedings of the ACM web conference 2022*, pages 2897–2905.

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

Wei-Lin Chiang, Zhuohan Li, Ziqing Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, and Joseph E. Gonzalez. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. *See https://vicuna. lmsys. org (accessed 14 April 2023)*, 2(3):6.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, and Evan Rosen. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

Alexey Dosovitskiy. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Wei Wei Du, Hong Wei Wu, Wei Yao Wang, and Wen Chih Peng. 2022. Team triple-check at factify 2: Parameter-efficient large foundation models with feature representations for multi-modal fact verification. In *CEUR Workshop Proceedings*, volume 3555. CEUR-WS.

Linmei Hu, Siqi Wei, Ziwang Zhao, and Bin Wu. 2022. Deep learning for fake news detection: A comprehensive survey. *AI open*, 3:133–155. ISBN: 2666-6510 Publisher: Elsevier.

Linmei Hu, Tianchi Yang, Luhao Zhang, Wanjun Zhong, Duyu Tang, Chuan Shi, Nan Duan, and Ming Zhou. 2021. Compare to the knowledge: Graph neural fake news detection with external knowledge. In *Proceedings of the 59th annual meeting of the association for computational linguistics and the 11th international joint conference on natural language processing (volume 1: long papers)*, pages 754–763.

Bo Li, Kaichen Zhang, Hao Zhang, Dong Guo, Renrui Zhang, Feng Li, Yuanhan Zhang, Ziwei Liu, and Chunyuan Li. 2024. LLaVA-NeXT: Stronger LLMs Supercharge Multimodal Capabilities in the Wild.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023a. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916.

Hui Liu, Wenya Wang, and Haoliang Li. 2023b. Interpretable multimodal misinformation detection with logic reasoning. In *61st Annual Meeting of the Association for Computational Linguistics (ACL 2023)*, pages 9781–9796. Association for Computational Linguistics.

Xuannan Liu, Peipei Li, Huaibo Huang, Zekun Li, Xing Cui, Jiahao Liang, Lixiong Qin, Weihong Deng, and Zhaofeng He. 2024. Fka-owl: Advancing multimodal fake news detection through knowledge-augmented lvlms. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 10154–10163.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

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.

Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Rose Finkel, Steven Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations*, pages 55–60.

Yida Mu, Xingyi Song, Kalina Bontcheva, and Nikolaos Aletras. 2024. Examining the limitations of computational rumor detection models trained on static datasets. In *Proceedings of the 2024 joint international conference on computational linguistics, language resources and evaluation (LREC-COLING 2024)*, pages 6739–6751.

Qiong Nan, Qiang Sheng, Juan Cao, Beizhe Hu, Danding Wang, and Jintao Li. 2024. Let silence speak: Enhancing fake news detection with generated comments from large language models. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, pages 1732–1742.

Hui Pang, Chaozhuo Li, Litian Zhang, Senzhang Wang, and Xi Zhang. 2025. Beyond Text: Fine-Grained Multi-Modal Fact Verification with Hypergraph Transformers. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 6389–6397. Issue: 6.

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

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.

Sahar Tahmasebi, Eric Müller-Budack, and Ralph Ewerth. 2024. Multimodal misinformation detection using large vision-language models. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, pages 2189–2199.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, and Faisal Azhar. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Pim Jordi Verschuuren, Jie Gao, Adelize van Eeden, Stylianos Oikonomou, and Anil Bandhakavi. 2023. Logically at Factify 2: A multi-modal fact checking system based on evidence retrieval techniques and transformer encoder architecture. *arXiv preprint arXiv:2301.03127*.

Herun Wan, Shangbin Feng, Zhaoxuan Tan, Heng Wang, Yulia Tsvetkov, and Minnan Luo. 2024. Dell: Generating reactions and explanations for llm-based misinformation detection. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 2637–2667.

Bo Wang, Jing Ma, Hongzhan Lin, Zhiwei Yang, Ruichao Yang, Yuan Tian, and Yi Chang. 2024a. Explainable fake news detection with large language model via defense among competing wisdom. In *Proceedings of the ACM Web Conference 2024*, pages 2452–2463.

Jiaqi Wang, Hanqi Jiang, Yiheng Liu, Chong Ma, Xu Zhang, Yi Pan, Mengyuan Liu, Peiran Gu, Sichen Xia, and Wenjun Li. 2024b. A comprehensive review of multimodal large language models: Performance and challenges across different tasks. *arXiv preprint arXiv:2408.01319*.

Jinguang Wang, Shengsheng Qian, Jun Hu, Wenxiang Dong, Xudong Huang, and Richang Hong. 2025. End-to-End Explainable Fake News Detection Via Evidence-Claim Variational Causal Inference. *ACM Transactions on Information Systems*, 43(4):1–26. ISBN: 1046-8188 Publisher: ACM New York, NY.

Youze Wang, Shengsheng Qian, Jun Hu, Quan Fang, and Changsheng Xu. 2020. Fake news detection via knowledge-driven multimodal graph convolutional networks. In *Proceedings of the 2020 international conference on multimedia retrieval*, pages 540–547.

Rongwu Xu, Brian Lin, Shujian Yang, Tianqi Zhang, Weiyan Shi, Tianwei Zhang, Zhixuan Fang, Wei Xu, and Han Qiu. 2024. The earth is flat because...: Investigating llms' belief towards misinformation via persuasive conversation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16259–16303.

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.

Wenjia Zhang, Lin Gui, and Yulan He. 2021. Supervised contrastive learning for multimodal unreliable news detection in COVID-19 pandemic. In *Proceedings of the 30th ACM international conference on information & knowledge management*, pages 3637–3641.

Yinuo Zhang, Zhulin Tao, Xi Wang, and Tongyue Wang. 2023. Ino at factify 2: Structure coherence based multi-modal fact verification. *arXiv preprint arXiv:2303.01510*.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, and Zican Dong. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223*, 1(2).

Xiaofan Zheng, Zinan Zeng, Heng Wang, Yuyang Bai, Yuhan Liu, and Minnan Luo. 2025. From predictions to analyses: Rationale-augmented fake news detection with large vision-language models. In *Proceedings of the ACM on Web Conference 2025*, pages 5364–5375.

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.

Xinyi Zhou, Jindi Wu, and Reza Zafarani. 2020. SAFE: Similarity-aware multi-modal fake news detection. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 354–367.

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

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2024. Minigpt-4: Enhancing vision-language understanding with advanced large language models. In *12th International Conference on Learning Representations, ICLR 2024*.

A Guiding Prompt Details

In guiding LVLMs to scrutinize news content, our strategy diverges from broad reasoning instructions. Instead, we enforce a structured decomposition of news narratives, explicitly directing the model's focus toward specific journalistic elements: time, location, subject-object dependencies, and causal chains. By constraining the model's attention to these atomic components, we facilitate a precise cross-modal alignment between the visual evidence, the claim's entities, and the authoritative text. While we acknowledge that the design space for prompting is vast and we do not claim our manually crafted templates represent the global optimum, our studies confirm that this element-centric logic significantly outperforms generic prompts in reducing hallucinations and clarifying the reasoning process. The specific prompt templates designed for the analysis and explanation generation stages are presented below.

A.1 Prompt for Analysis Generation

"<*image evidence*> Claim text: <*news claim*> Entities in claim: <*entities*> Evidence retrieved from authoritative news sources: <*text evidence*> Please comprehensively analyze the claim text and image, focusing on subject-object relationships, time and location information, and causal logic. Discuss whether the described interactions, time/place, and cause-effect are reasonable and consistent with the evidence. These clues may be used with others to further predict the truth of the news, so your answer does not need to provide a definitive conclusion. Limit your response to 30 words."

A.2 Prompt for Explanation Generation

"<*image evidence*> Claim text: <*news claim*> Entities in claim: <*entities*> Evidence retrieved from authoritative news sources: <*text evidence*> Analysis: <*analysis*> This is a preliminary analysis of claim before determining its truthfulness. Now the claim is determined to be <*refuted / supported / not enough information*>. Please provide a detailed explanation of why the claim is <*refuted / supported / not enough information*>, based on the analysis above. Limit your response to 50 words."

B Examples of Generated Explanations

To complement the NEI example discussed in the main text (see Figure 1), we provide additional qualitative case studies in Figure 4 to further elucidate our model's performance on claims with definitive binary veracity labels. Specifically, we present a detailed visualization of one refuted case and one supported case to demonstrate the model's reasoning versatility: for the refuted claim, HPIM successfully detects semantic contradictions to generate a corrective justification, while for the supported claim, it aggregates corroborating details to validate the statement. These examples collectively highlight how HPIM effectively attends to and synthesizes critical evidence to substantiate its predictions when sufficient information is present in the source documents.

News Claim: Refuted
Facebook users are entitled to $17,500 each as compensation over a 'data breach' involving Cambridge Analytica.

Text Evidence:
All Facebook Users Could Cash in as much $17,500 Each After Data Breach. If your data was harvested through Facebook you could get £12,500 compensation, according to an expert. The social network has come under fire after it was revealed Cambridge Analytica kept users' data. This could cost Facebook £625 billion, which is double the £317b it is worth, law professor Maureen Mapp argued. 'There are about 50 million users whose data was harvested,' she told the Sun. 'Assuming each one of them brought a claim for compensation for distress caused by the data breach ... each individual may be awarded £12,500 as damages ... But a more likely outcome is that users would receive a maximum of £500 each, according to data protection lawyer David Barda, who works for Slater and Gordon. He added: 'The amount of compensation will depend on the level of distress suffered, but Facebook could be facing claims of up to £500 per Facebook user if those users were able to demonstrate their distress.'

Explanation:
The claim of Facebook users being entitled to $17,500 each as compensation for a data breach involving Cambridge Analytica is refuted because it is a misrepresentation of a single comment made by lawyer David Barda, and is not widely recognized or reported. Facebook users are not entitled to this amount or any compensation. The image features a Facebook auction symbolized by a dollar bill and a stack of 100 dollar bills, which suggests the idea of monetary compensation for users, but there is no factual or legal basis to support this claim.

News Claim: Supported
A map displayed during Sean Hannity's Fox News show mislabeled the Upper Peninsula of Michigan as Canada.

Text Evidence:
The Upper Peninsula is a lot of things: Beautiful, natural, rural, teeming with minerals, you name it. But it is most definitely and inarguably not Canada. Unfortunately, it was labeled as such during Thursday night's 'Hannity' on Fox News, when conservative talk show personality Sean Hannity was discussing Wisconsin's 2020 election recount efforts.

Explanation:
The claim that the error in labeling the Upper Peninsula of Michigan as "Canada" is unintentional misinformation is supported by several points. Firstly, the scenario is described as incidental rather than intentional, which suggests a lack of malice or deception. Secondly, the mislabeled map is described as a product of a news television show, which typically employs visual design priorities that may not always align with complete geographical accuracy. Finally, there are no explicit accusations of dishonesty or manipulation in the given context, supporting the conclusion that the error in labeling the Upper Peninsula is likely accidental rather than deliberate.

Figure 4: Qualitative examples of generated explanations for refuted and supported claims.