MUSE: Uncertainty-aware Modular Semantic Reasoning for Multimodal Incorrect Information Detection

Anonymous ACL submission

Abstract

Incorrect information can undermine content quality, veracity, and integrity, and these challenges are strengthened when inaccurate information is shared within multimodal information streams. Multimodal detection methods usually fuse textual, visual, or external knowledge features to detect incorrect information, but ignore (i) reasoning uncertainty across modalities, (ii) contrastive interplay between internal content verification and external knowledge modification, and (iii) interpretable diagnostics of why and when a reasoning head changes its prediction when exposed to auxiliary knowledge, including external knowledge and visual modality. To address these issues, we propose a Multimodal Uncertainty-aware Semantic rEasoning framework (MUSE) that disentangles content-internal reasoning from knowledge-augmented reasoning. First, MUSE decomposes multimodal reasoning into modular reasoning heads. Then, we calibrate the predictions from modular reasoning heads via an uncertainty-aware fusion mechanism, enabling the model to adaptively balance content signals and auxiliary knowledge under varying uncertainty conditions. To quantify the contributions of auxiliary knowledge, we further define a novel metric, the Value of eXperience (VoX), to measure sample-level logit gains from knowledge or vision reasoning. Importantly, MUSE is backbone-agnostic and can be integrated with multiple multimodal encoders. Experimental results demonstrate that MUSE achieves optimal accuracy, providing a transparent, modular reasoning process and interpretable diagnostics on when and why knowledge matters.

1 Introduction

As individuals and communities increasingly turn to online social media platforms like X (Twitter), Instagram, and TikTok for communication, interaction, or education, we must consider the kinds of information people encounter in these spaces. Though these platforms offer convenience and accessibility, the factual accuracy of digital content often falls short of media standards and societal expectations (Shu et al., 2017; Zhou and Zafarani, 2020; Abdali et al., 2024). A substantial body of unverified and distorted content is produced and circulated on social media platforms (Su et al., 2020; Ngueajio et al., 2025). Such incorrect and misleading information (e.g., spam, rumors, etc.) poses significant challenges to individuals and society, undermining public trust and media credibility, as well as contributing to suboptimal health beliefs and decisions (Thorson et al., 2010; Bhattarai et al., 2021; Schmid and Bauer, 2025). Such relationships are further complicated as multimodal incorrect information, including textual, visual, or aural elements, has become increasingly prevalent on social media, such as memes (Zhong and Baghel, 2024) and deepfakes (Huang et al., 2025), undermining the digital ecosystem. To address these

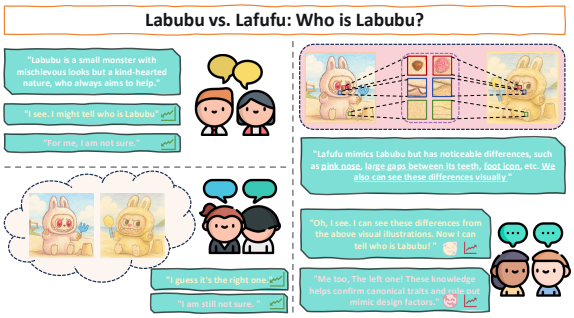

Figure 1: **Illustration of contrastive reasoning across textual, visual, and knowledge heads.** Top-left: without auxiliary knowledge, the content-internal head has low confidence ↘ and cannot tell Labubu. Bottom-left: with visual signals, the visual reasoning head labels Lafufu (counterfeit Labubu) as Labubu with low certainty ↘. Right: with textual, visual, and external knowledge being jointly integrated, the knowledge-augmented head corrects the label as Lafufu with high confidence ↗.

challenges, automatic models, involving unimodal and multimodal frameworks, are proposed to detect and filter the incorrect information, mitigating the widespread effects of incorrect information on social media, mass communication, and public health (Guo et al., 2020; Yang et al., 2023; Shi et al., 2023).

Although information detection methods have achieved notable improvements, these methods still struggle with disentangling complex multimodal features, integrating auxiliary knowledge in a reliable manner, and mitigating modality uncertainty and bias. Most approaches focus primarily on either textual content features (e.g., linguistic features and semantic embeddings) or on auxiliary knowledge (e.g., visual modality, knowledge bank, and retrieval-augmented generation). But, these methods ignore cross-modality uncertainty and integrate all feature streams into a single classifier, without any disentanglement between what is learned from textual content and what is contributed by auxiliary knowledge. As shown in Figure 1, it illustrates how knowledge helps form more accurate beliefs when facing potentially misleading content. Without relevant knowledge inputs, the reader accepts the incorrect information (i.e., Lafufu) as correct (i.e., Labubu) due to their similar names and visual appearances. But once the reader has relevant knowledge about Labubu's canonical features, the information is recognized as incorrect.

To uncover interplay between content-internal reasoning and knowledge-augmented reasoning (i.e., visual inputs or external knowledge), we propose a new view: disentangling content reasoning from knowledge reasoning in a modular way. In this paper, we introduce a novel multimodal modular-head model architecture, **MUSE**, for incorrect information detection. Our framework employs contrastive and modular heads to explore the textual, visual, and knowledge features, respectively, where the knowledge is retrieved from an in-domain memory or external knowledge source. This separation mechanism allows us to measure, analyze, and apply these feature streams in a structured and quantitative way. Technically, our core contributions are summarized as follows:

- First, we introduce a multimodal, modular, and contrastive semantic reasoning framework that disentangles content-internal reasoning from knowledge-augmented reasoning.
- Second, we propose an uncertainty-aware multimodal fusion mechanism with cross-head refinement to adaptively incorporate textual, visual, and knowledge features, where fusion weights are calibrated using intra-branch uncertainty, inter-branch disagreement, and retrieval uncertainty.
- Third, we define a Value-of-eXperience (VoX) metric to quantify the contributions of auxiliary knowledge and unfold when and why knowledge matters, improving model interpretability.

Experimental results on Twitter and Weibo datasets demonstrate that MUSE improves detection accuracy and interpretability, especially when the auxiliary knowledge contributes to predictions.

2 Related work

2.1 Content-internal reasoning methods

Traditional machine learning (ML) methods have been employed to classify the textual content of information units into predefined classes; however, these models typically rely on hand-crafted features and struggle to capture complex semantic patterns, thus limiting detection accuracy (Minaee et al., 2021). As neural networks have emerged, deep learning frameworks have further improved performance by extracting complex content features and capturing nuanced semantic representations, such as convolutional neural networks (Kim, 2014; Li et al., 2017; Wang, 2017; Kaliyar et al., 2020), recurrent neural networks (Wang et al., 2016; Ma et al., 2016), and long short-term memory (Yue-Hei Ng et al., 2015; Sachan et al., 2019; Ma et al., 2020).

In addition, attention networks improve model performance by incorporating feature embeddings extracted from different dimensions of information units (Yang et al., 2016; Linmei et al., 2019; Sun and Lu, 2020; Lovisotto et al., 2022). For instance, Yang et al. (Yang et al., 2016) explored the word-level and sentence-level attentions and presented a hierarchical attention network to capture document hierarchy. For understanding visual signals, Fukui et al. (Fukui et al., 2019) proposed an attention branch network to incorporate visual explanations with attentions, improving model interpretability. By introducing pre-trained embeddings, the transformer-based methods have further boosted detection accuracy for incorrect information content (Croce et al., 2020; Xiong et al., 2021;

Van Nooten and Daelemans, 2025). Wu et al. (Wu et al., 2024) emphasized the substantive content over stylistic features, using Large Language Models (LLMs) to reframe news articles and focus on content correctness.

2.2 Knowledge-augmented reasoning methods

As data structures and modalities continue to expand, multimodal approaches (Yuan et al., 2021; Wang et al., 2022; Lin et al., 2023) have been proposed to handle more intricate information reasoning tasks by incorporating external knowledge sources and reasoning across text, images, video, audio data, or multiple languages (Conneau and Lample, 2019; Wang et al., 2023; Abdali et al., 2024; Wu et al., 2024; Ganz et al., 2024; Li et al., 2025). For instance, Wang et al. (Wang et al., 2018) proposed event adversarial neural networks (EANN) to detect multimodal incorrect information, where EANN combines the event discriminator into the detection model, learning event features from text and image data. Though it has capacity to detect unseen topics or events, EANN relies heavily on the quality of event features. Khattar et al. (Khattar et al., 2019) presented a multimodal variational autoencoder (MVAE) model, to learn a shared representation of textual and visual content within an information unit and capture cross-modality dependency between texts and images. However, MVAE can't handle the noise or information loss in the generative process, and also has no discrimination before feature fusion.

Additionally, Zhou et al. (Zhou et al., 2020) proposed a similarity-aware multimodal false news detection (SAFE) method. SAFE explores the similarity between textual and visual representations and ensures the consistency between text and image. However, it only identifies one type of inconsistency, and similarity evaluation can't capture the complex dependency between text and image data. To evaluate the ambiguity between textual and visual data, Chen et al. (Chen et al., 2022) proposed the cross-modal ambiguity learning for incorrect information detection (CAFE). CAFE introduced an adaptive feature fusion scheme using an ambiguity-aware alignment module, where it can adaptively weight text and image representations by measuring the ambiguity. Though achieving better performance, it ignores the contextual knowledge of multimodal information and does not incorporate uncertainty and consistency measurements, where it does not quantify the contributions from different modalities and only feeds calibrated features into the classifier.

Another line of incorrect information detection research focuses on enhancing detection performance through knowledge augmentation. For example, retrieval-augmented methods (Guu et al., 2020; Lewis et al., 2020) incorporate the retrieval knowledge units into LLMs to improve factual reasoning. Recent work introduced a dual-head reasoning model (Zhang et al., 2026) for textual content detection. However, it's limited to textual data and does not incorporate multimodal inputs, such as visual signals. Cheng et al. (Cheng et al., 2025) introduced a question-answering benchmark dataset for factuality evaluation in multimodal LLMs. Though LLMs emerged with impressive capability of processing multimodal features, LLMs still require a large volume of data to update the known knowledge and maintain performance and reliability (Lin et al., 2025; Fang et al., 2025). Currently, vision-language models (VLMs) (Dai et al., 2023; Liu et al., 2023; Bai et al., 2025) have shown strong potential for multimodal reasoning by jointly encoding textual and visual information and generating semantic representations across modalities, where VLMs leverage large-scale pretraining to capture visual semantics, textual context, and cross-modal associations. However, VLMs still suffer from outdated or incomplete knowledge, hallucinated reasoning, and insufficient uncertainty calibration when handling noisy multimodal evidence.

In conclusion, the existing multimodal detection models typically combine textual, visual, and knowledge features into a single semantic representation. These methods have two key limitations. First, it hinders understanding of how auxiliary knowledge shifts predictions. Second, multimodal features are inherently heterogeneous, differing in noise distribution and semantic representation; however, conventional fusion strategies integrate these streams in an uncertainty-agnostic manner, potentially amplifying noisy or unreliable signals. MUSE addresses these limitations through a modular reasoning framework and the uncertainty-aware fusion mechanism, which dynamically calibrates predictions from multiple reasoning heads.

3 Methodology

In this section, we introduce the multimodal framework of our MUSE, as shown in Figure 2. Our objective is to disentangle content-internal reasoning

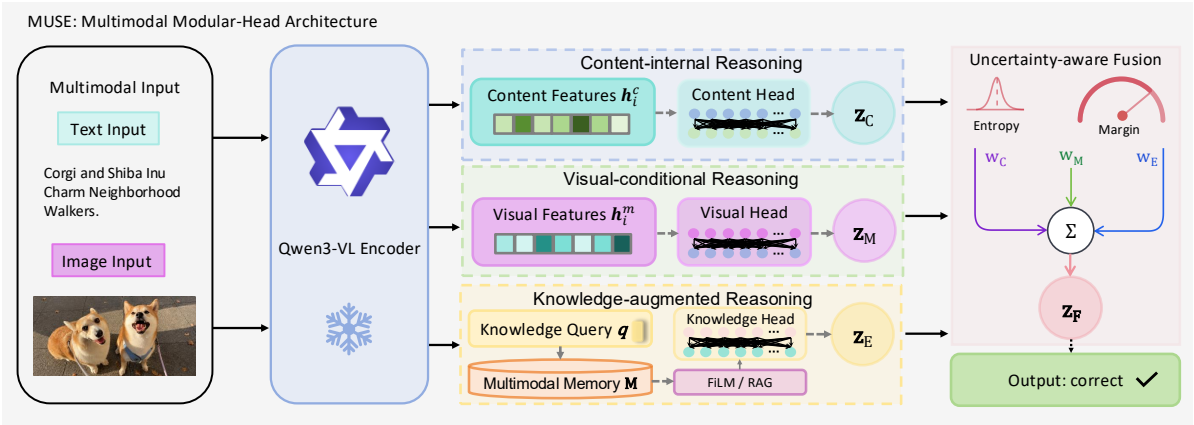

Figure 2: **An illustration of the proposed MUSE framework.** (a) **Multimodal Modular-head Architecture:** Input information unit is first encoded via Qwen3-VL model. Then, the *content-internal head* performs text-centered reasoning and produces logits $\mathbf{z}_{\mathrm{C}}$, and the *visual-conditioned head* captures cross-modality reasoning and outputs logits $\mathbf{z}_{\mathrm{M}}$. (b) In parallel, the **knowledge-augmented branch** retrieves relevant neighbors from a multimodal memory bank and injects the retrieved knowledge via FiLM/RAG, generating knowledge-enhanced logits $\mathbf{z}_{\mathrm{E}}$. (c) **Uncertainty-aware Fusion:** Multimodal features are integrated by an uncertainty-aware fusion strategy, producing final logits $\mathbf{z}_{\mathrm{F}}$.

from knowledge-augmented reasoning in a modular manner and provide interpretable diagnostics on when and why auxiliary knowledge, including visual and retrieved knowledge, shifts the predictions. In the MUSE paradigm, the core modules are as follows: (i) a multimodal encoder; (ii) a content-internal reasoning module; (iii) a visual/knowledge-augmented reasoning module; (iv) an uncertainty-aware fusion mechanism, where MUSE allows the content reasoning to dynamically interact with and refine the knowledge branch before final fusion; and (v) a quantitative VoX metric at the instance level. Importantly, the MUSE framework is backbone-agnostic and can be integrated with multiple pretrained encoders.

3.1 Correctness-oriented problem definition

Information units can be decomposed into a set of semantic elements, such as entities, events, sentiments, behaviors, etc. These elements are usually conveyed by textual content, visual modality, or contextual interactions. In this paper, we introduce a correctness-oriented definition at the information-unit level and transform incorrect information detection into assessing whether an information unit x_i is correct. Our assessment foundation is that x_i is correct if it contains no incorrect components in the incorrectness set $\mathcal{I}(x_i)$, such as textual and visual elements. Therefore, the information detection task is reframed as identifying incorrect components within x_i. Here, we formalize the assessment system as:

$$y(x_i) = \begin{cases} 1 & \text{if } \mathcal{I}(x_i) = \varnothing \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

Our objective is to train a classifier on multimodal information units:

$$f(x_i; \boldsymbol{\theta}) = \hat{y}_i \tag{2}$$

in which $\hat{y}_i \in \{0, 1\}$ denotes the predicted label of correctness and $\boldsymbol{\theta}$ is the learnable model parameters.

3.2 Multimodal encoder

Instead of designing separate textual and visual encoders with multimodal interaction modules, we employ a single frozen Qwen3-VL-2B (Bai et al., 2025) backbone to extract three complementary branch representations from input data. Given an information unit x_i, including text T_i and image I_i, we construct three prompt variants: a textual prompt for content reasoning, a visual prompt for visual reasoning, and a text-image prompt for multimodal reasoning. Each prompt is processed by the Qwen3-VL processor and passed through the same frozen multimodal backbone. After tokenization and multimodal encoding, the Qwen3-VL model $\boldsymbol{E}_{\text{Qwen-VL}}(\cdot)$ produces final-layer hidden states:

$$\begin{aligned} \boldsymbol{H}_i^c &= \boldsymbol{E}_{\text{Qwen-VL}}(T_i), \\ \boldsymbol{H}_i^v &= \boldsymbol{E}_{\text{Qwen-VL}}(I_i), \\ \boldsymbol{H}_i^m &= \boldsymbol{E}_{\text{Qwen-VL}}(T_i, I_i), \end{aligned} \tag{3}$$

where $\boldsymbol{H}_i^c \in \mathbb{R}^{L_c \times d}$, $\boldsymbol{H}_i^v \in \mathbb{R}^{L_v \times d}$, and $\boldsymbol{H}_i^m \in \mathbb{R}^{L_m \times d}$ are hidden states from the content/text branch, visual branch, and multimodal branch, respectively. Here, L_c, L_v, and L_m are the sequence lengths of the three prompt variants, and d is the hidden dimension of the Qwen3-VL model.

In our framework, the representation of each branch is produced from the last hidden layer:

$$\begin{aligned} \boldsymbol{h}_i^c &= \boldsymbol{H}_i^c[L_c, :], \\ \boldsymbol{h}_i^v &= \boldsymbol{H}_i^v[L_v, :], \\ \boldsymbol{h}_i^m &= \boldsymbol{H}_i^m[L_m, :]. \end{aligned} \tag{4}$$

$\boldsymbol{h}_i^c$ denotes the content feature extracted, $\boldsymbol{h}_i^v$ denotes the visual feature, and $\boldsymbol{h}_i^m$ denotes the multimodal feature. These three feature streams provide complementary reasoning evidence for incorrect information detection: the content branch focuses on linguistic reasoning, the visual branch performs visual-semantic reasoning, and the multimodal branch captures multimodal interactions. Notably, all three semantic representations are generated by the same frozen Qwen3-VL backbone under different prompting conditions.

3.3 Multimodal retrieval via memory bank

In this section, we construct a multimodal knowledge bank $\mathbf{M} = \{\boldsymbol{m}_1, \boldsymbol{m}_2, \ldots, \boldsymbol{m}_N\}$ for textual and visual data, respectively. $\boldsymbol{m}_j \in \mathbb{R}^{d_s}$ denotes the text or image embedding. d_s is the embedding dimension, which varies across modalities. N is the number of training units in $\mathbf{M}$. For an input x_i, we encode it as query $\boldsymbol{q} \in \mathbb{R}^{d_s}$ and compute cosine similarity s_i between $\boldsymbol{q}$ and $\boldsymbol{m}_j$. Then, we select the top-k nearest neighbors with indices $\mathcal{N}(x_i)$ and aggregate:

$$\bar{\boldsymbol{m}} = \frac{1}{k} \sum_{j \in \mathcal{N}(x_i)} \boldsymbol{m}_j. \tag{5}$$

To inject the retrieved knowledge neighbors, we transform $\bar{\boldsymbol{m}}$ into modulation parameters $\boldsymbol{\gamma}_k$ and $\boldsymbol{\beta}_k$:

$$\boldsymbol{\gamma}_k = \mathbf{W}_\gamma \bar{\boldsymbol{m}} + \mathbf{b}_\gamma, \quad \boldsymbol{\beta}_k = \mathbf{W}_\beta \bar{\boldsymbol{m}} + \mathbf{b}_\beta \tag{6}$$

where $\mathbf{W}_\gamma, \mathbf{W}_\beta \in \mathbb{R}^{d \times d_s}$ are projection matrices; $\mathbf{b}_\gamma$, $\mathbf{b}_\beta$ are bias terms. We then apply FiLM (Perez et al., 2018) to generate the knowledge-augmented representations $\boldsymbol{h}_{\mathrm{E}}$ for textual and visual embeddings, respectively. Here, $\odot$ is element-wise multiplication:

$$\boldsymbol{h}_{\mathrm{E}}^c = \boldsymbol{h}_i^c \odot (1 + \tanh(\boldsymbol{\gamma}_k)) + \boldsymbol{\beta}_k \tag{7}$$

In addition, we further explore retrieval-augmented VLMs to incorporate external knowledge sources into multimodal reasoning. More implementation details are provided in Suppl.

3.4 Multimodal reasoning with modular heads

Combining text representations with raw content features $\boldsymbol{c} \in \mathbb{R}^{d_c}$, we define two textual reasoning heads:

$$\mathbf{z}_{\mathrm{C}} = f(\boldsymbol{h}_{\boldsymbol{i}}^{\boldsymbol{c}}, \boldsymbol{c}; \boldsymbol{\theta}_{\mathrm{C}}), \qquad \mathbf{z}_{\mathrm{E}} = f(\boldsymbol{h}_{\mathrm{E}}^c, \boldsymbol{c}; \boldsymbol{\theta}_{\mathrm{E}}) \tag{8}$$

where $\mathbf{z}_{\mathrm{C}}, \mathbf{z}_{\mathrm{E}} \in \mathbb{R}^K$ are outputs of the textual content and knowledge-augmented reasoning heads, respectively. K is the number of labels. Then, $\mathbf{z}_{\mathrm{C}}$, $\mathbf{z}_{\mathrm{E}}$ are transformed into probability distributions through softmax:

$$\hat{\boldsymbol{y}}_{\mathrm{C}} = \mathrm{softmax}(\mathbf{z}_{\mathrm{C}}), \quad \hat{\boldsymbol{y}}_{\mathrm{E}} = \mathrm{softmax}(\mathbf{z}_{\mathrm{E}}) \tag{9}$$

$\hat{\boldsymbol{y}}_{\mathrm{C}}$ is the prediction results from content-internal reasoning $\mathrm{Head}_{\mathrm{C}}$, and it's what the model "believes" without external knowledge. $\mathrm{Head}_{\mathrm{E}}$ reflects knowledge-augmented reasoning and derives $\hat{\boldsymbol{y}}_{\mathrm{E}}$, and $\hat{\boldsymbol{y}}_{\mathrm{E}}$ is what the model "believes" with retrieved knowledge. Together, we can explore how knowledge affects predictions. More details are provided in Suppl.

3.5 Uncertainty-aware fusion strategy

To fuse the final predictions from modular heads, we introduce an uncertainty-aware fusion strategy. Here, we first define $\mathcal{H}$ as the set of heads $\{\mathrm{Head}_{\mathrm{C}}, \mathrm{Head}_{\mathrm{M}}, \mathrm{Head}_{\mathrm{E}}\}$ and compute uncertainty $\mathcal{U}_h$ for each head $h \in \mathcal{H}$:

$$\mathcal{U}_h = H\left(\frac{1}{S} \sum_{s=1}^{S} \mathrm{softmax}\left(\mathbf{z}_h^s\right)\right) \tag{10}$$

where $H(\cdot)$ is Shannon entropy function, S is the number of stochastic forward passes. Then, we define prediction confidence as the decision margin:

$$\mathcal{M}_h = p_h^1 - p_h^2 \tag{11}$$

Here, p_h^1 and p_h^2 denote the largest and second-largest probabilities from the predictive distribution of head h. Furthermore, we formulate the gate vector $\mathbf{u}$, using $\mathcal{U}_h$ and $\mathcal{M}_h$, and feed it into a multi-layer perceptron to produce the weight vector $\mathbf{w} = \{\mathrm{w}_{\mathrm{C}}, \mathrm{w}_{\mathrm{M}}, \mathrm{w}_{\mathrm{E}}\}$, where $\mathrm{w}_k >= 0$ and $\sum_k \mathrm{w}_k = 1$.

Finally, we fuse logits:

$$\mathbf{z}_{\mathrm{F}} = \sum_{k \in \{\mathrm{C,M,E}\}} \mathrm{w}_k \cdot \mathbf{z}_k, \quad \hat{\boldsymbol{y}}_{\mathrm{F}} = \mathrm{softmax}(\mathbf{z}_{\mathrm{F}}) \tag{12}$$

where $\mathbf{z}_{\mathrm{F}}$ is the fused logits combining content, visual, and knowledge reasoning heads. $\hat{\boldsymbol{y}}_{\mathrm{F}}$ is the final prediction results. More details can be found in Suppl.

3.6 Agreement regularization

To stabilize training and enforce agreement, we define the symmetric KL regularizer as:

$$\mathcal{L}_{\text{agree}}^{(i)} = \tfrac{1}{2}\left[D_{\mathrm{KL}}(\mathbf{z}_{\mathrm{C}}^{(i)} \,\|\, \mathbf{z}_{\mathrm{E}}^{(i)}) + D_{\mathrm{KL}}(\mathbf{z}_{\mathrm{E}}^{(i)} \,\|\, \mathbf{z}_{\mathrm{C}}^{(i)})\right] \tag{13}$$

where $D_{\mathrm{KL}}(\mathbf{z}_{\mathrm{C}}^{(i)} \,\|\, \mathbf{z}_{\mathrm{E}}^{(i)})$ is the KL divergence between distributions $\mathbf{z}_{\mathrm{C}}^{(i)}$ and $\mathbf{z}_{\mathrm{E}}^{(i)}$. $\mathcal{L}_{\text{agree}}^{(i)}$ is agreement loss. Then, we introduce auxiliary branch supervision loss $\mathcal{L}_{\text{heads}}^{(i)}$. Finally, our training objective is:

$$\mathcal{L} = \frac{1}{n}\sum_{i=1}^{n}\Big[\mathcal{L}_{\mathrm{CE}}(\boldsymbol{z}_{\mathrm{F}}^{(i)}, y(x_i)) + \lambda_{agree}\,\mathcal{L}_{\text{agree}}^{(i)} + \lambda_{heads}\,\mathcal{L}_{\text{heads}}^{(i)}\Big] \tag{14}$$

where n is the number of training data. $\lambda_{agree} = 0.05$ and $\lambda_{heads} = 0.3$ are weight parameters. Through the loss function $\mathcal{L}$, we encourage both heads to produce consistent predictions while maintaining their distinct reasoning. In addition, we employ the Adam optimizer, mixed-precision training, and early stopping to effectively reach convergence. More method details are provided in Suppl.

3.7 Value-of-eXperience metric

To measure the contributions of auxiliary knowledge, we define VoX at the instance level. Given the correctness label $y(x_i)$, VoX is:

$$\mathrm{VoX}(x_i) = \mathbf{z}_{\mathrm{F}}[y(x_i)] - \mathbf{z}_{\mathrm{C}}[y(x_i)]. \tag{15}$$

where $\mathbf{z}_{\mathrm{F}}$ is the final output integrating the modular heads. We further decompose the multimodal reasoning into per-head contributions via VoX, such as $\mathbf{z}_{\mathrm{E}}[y(x_i)] - \mathbf{z}_{\mathrm{C}}[y(x_i)]$. Our evaluation criteria are listed as follows:

- $\mathrm{VoX}(x_i) > 0$: knowledge increases confidence in correct class.
- $\mathrm{VoX}(x_i) < 0$: knowledge decreases confidence, suggesting potential noise.
- $\mathrm{VoX}(x_i) \approx 0$: knowledge has little effect.

We report the mean, median, and proportion of positive-VoX samples to provide interpretable diagnostics on knowledge augmentation. Beyond raw accuracy or F1 score, VoX discloses *when and why* knowledge matters in an explainable, modular reasoning framework.

4 Experiments

In this section, we conducted comparison experiments on multimodal benchmark datasets, and the experimental results demonstrate that MUSE has superior performance to most tested models. We first introduced the experimental setup, including the datasets, tested models, and experiment settings. Then, we reported the experiment results and VoX values, and then analyzed these results for further exploration. Furthermore, the ablation study shows the modules contributing to the performance improvement.

4.1 Experimental Setup

Datasets. For conducting comparison experiments, we test our MUSE and other advanced models on Twitter (Boididou et al., 2015) and Weibo (Jin et al., 2017) datasets.

Experimental Models. To perform the comparison experiments fairly, we compared our proposed MUSE model with 9 tested models, which include EANN (Wang et al., 2018), MVAE (Khattar et al., 2019), MKEMN (Zhang et al., 2019), SAFE (Zhou et al., 2020), MCNN (Xue et al., 2021), MCAN (Wu et al., 2021), CAFE (Chen et al., 2022), FND-CLIP (Zhou et al., 2023), and Event-Radar (Ma et al., 2024). More experimental details are provided in Suppl.

4.2 Experimental results

In Table 1, we reported the experimental results of all the tested models across these two datasets. As shown in Table 1, one can see that our MUSE model achieves superior performance across most metrics on both the Twitter and Weibo datasets, such as Acc, Rec, and F1. It shows that our proposed MUSE model can incorporate textual features, visual signals, and external knowledge in an efficient and structured way, where MUSE captures the cross-modality uncertainty and learns the contrastive interplay between internal content verification and external knowledge modification. Additionally, we can see that our model achieves the

Method	Twitter Dataset						
	Acc	Incorrect Information Unit			Correct Information Unit		
		Pre	Rec	F1	Pre	Rec	F1
EANN (Wang et al., 2018)	0.648	0.810	0.498	0.617	0.584	0.759	0.660
MVAE (Khattar et al., 2019)	0.745	0.801	0.719	0.758	0.689	0.777	0.730
MKEMN (Zhang et al., 2019)	0.715	0.814	0.756	0.708	0.634	0.774	0.660
SAFE (Zhou et al., 2020)	0.762	0.831	0.724	0.774	0.695	0.811	0.748
MCNN (Xue et al., 2021)	0.784	0.778	0.781	0.779	0.790	0.787	0.788
MCAN (Wu et al., 2021)	0.809	**0.889**	0.765	0.822	0.732	**0.871**	0.795
CAFE (Chen et al., 2022)	0.806	0.807	0.799	0.803	0.805	0.813	**0.809**
FND-CLIP (Zhou et al., 2023)	0.725	0.701	0.904	0.790	0.792	0.487	0.603
Event-Radar (Ma et al., 2024)	0.702	0.716	0.795	0.753	0.678	0.579	0.624
MUSE	**0.831**	0.814	**0.913**	**0.861**	**0.862**	0.722	0.768
	Weibo Dataset						
EANN (Wang et al., 2018)	0.827	0.847	0.812	0.829	0.807	0.843	0.825
MVAE (Khattar et al., 2019)	0.824	0.854	0.769	0.809	0.802	0.875	0.837
MKEMN (Zhang et al., 2019)	0.814	0.823	0.799	0.812	0.723	0.819	0.798
SAFE (Zhou et al., 2020)	0.816	0.818	0.815	0.817	0.816	0.818	0.817
MCNN (Xue et al., 2021)	0.823	0.858	0.801	0.828	0.787	0.848	0.816
MCAN (Wu et al., 2021)	0.829	0.589	0.889	0.711	**0.970**	0.809	0.887
CAFE (Chen et al., 2022)	0.840	0.855	0.830	0.842	0.825	0.851	0.837
FND-CLIP (Zhou et al., 2023)	0.889	0.874	0.894	0.884	0.904	0.885	0.895
Event-Radar (Ma et al., 2024)	0.907	0.914	0.901	0.908	0.914	0.901	0.907
MUSE	**0.962**	**0.957**	**0.969**	**0.963**	0.969	**0.956**	**0.962**

Table 1: Performance comparison on Twitter and Weibo datasets, including accuracy (Acc), precision (Pre), recall (Rec), and F1-score (F1). The best results are highlighted in **bold**.

highest recall value on both Twitter and Weibo, especially for the incorrect information unit. A higher recall indicates that less incorrect information is missed. Furthermore, it should be noted that our model maintains robust and consistent performance across both datasets, achieving more balanced Pre and Rec values compared with other tested models.

For the comparison models, EANN (Wang et al., 2018), MVAE (Khattar et al., 2019), and MKEMN (Zhang et al., 2019) show moderate performance on both datasets, indicating the limited capacity to effectively integrate multimodal features. One can see that SAFE (Zhou et al., 2020) and MCNN (Xue et al., 2021) present improved performance. Notably, MCAN (Wu et al., 2021) and CAFE (Chen et al., 2022) receive competitive results across evaluation metrics on both datasets, suggesting that these models can capture the interactions between textual and visual modalities. Although FND-CLIP (Zhou et al., 2023) and Event-Radar (Ma et al., 2024) achieve higher accuracy on Weibo compared with earlier methods, their performance drops significantly on Twitter, indicating limited robustness and generalization across datasets. In conclusion, our MUSE model demonstrates strong capacity in detecting incorrect information and provides more balanced performance across Acc, Pre, Rec, and F1. Additionally, MUSE shows stable performance across runs with low variance on both datasets.

4.3 VoX analysis

From Table 2, the VoX results vary with different modular head settings. For example, on the Weibo dataset, the Head_M produces expected positive VoX gains (+1.40 / 90.28%), and Head_E also presents a positive effect (+2.04 / 90.13%) on shifting predictions, implying that multimodal and retrieved knowledge can enhance content-internal reasoning and improve predictions. In contrast, the final fused setting $\text{Head}_\text{C} \oplus \text{Head}_\text{M} \oplus \text{Head}_\text{E}$ achieves a lower VoX gain (+1.16 / 90.93%), indicating that integrating knowledge does not always help multimodal reasoning to learn high-quality feature representations. Although the final fused model has a smaller VoX magnitude (+1.16), it achieves the highest positive VoX ratio (90.93%), showing that uncertainty-aware fusion stabilizes prediction updates and reduces over-aggressive reasoning shifts. Additionally, compared Head_M and Head_E, it is worth noting that stronger prediction shifts (+1.40 VoX vs. +2.04 VoX) do not always translate into better final decisions; robust multimodal reasoning may outperform knowledge intervention. More

Dataset	Head/Mode	Acc	F1	Pre	Rec	VoX (mean / pos%)
Weibo	Head_C	85.72	84.99	89.77	80.70	–
	Head_M	96.14	96.10	97.33	94.90	+1.40 / 90.28%
	Head_E	95.49	95.43	97.00	93.90	+2.04 / 90.13%
	$\text{Head}_\text{C} \oplus \text{Head}_\text{M} \oplus \text{Head}_\text{E}$	95.94	95.96	95.63	96.30	+1.16 / 90.93%

Table 2: Head-wise ablation results on the Weibo dataset. Head_C denotes the content-only branch, Head_M denotes the multimodal branch, and Head_E denotes the knowledge-augmented branch. VoX is computed relative to Head_C.

Variant	Acc	F1	Macro-F1	VoX
Full MUSE	0.9534	0.9542	0.9534	0.7983
w/o **C** pathway	0.9519	0.9526	0.9519	–
w/o **M** pathway	0.9494	0.9509	0.9493	1.0038
w/o **E** pathway	0.9589	0.9590	0.9589	0.8405
w/o Uncertainty Fusion	0.9509	0.9521	0.9509	0.9197

Table 3: Ablation study on the Weibo dataset, where content, visual, and knowledge are denoted as **C**, **M**, and **E**, respectively.

detailed analysis is provided in Suppl.

4.4 Ablation study

As shown in Table 3, the full MUSE model achieves strong performance on the Weibo dataset (Acc/F1 = 0.9534/0.9542). Removing the content pathway, multimodal pathway, or uncertainty fusion leads to performance degradation, showing that these components provide complementary contributions to the overall effectiveness of the MUSE framework.

On the Weibo dataset, when removing the visual signals (i.e., **M**), it yields the largest performance drop (Acc 0.9534 $\rightarrow$ 0.9494), indicating that visual signals have a significant impact on Weibo information units where images play a crucial role in Weibo posts. Removing the content pathway (i.e., **C**) also reduces accuracy (0.9534 $\rightarrow$ 0.9519), showing that textual content remains an important source of evidence. Interestingly, the knowledge pathway (i.e., **E**) degrades performance without uncertainty-aware fusion, dropping accuracy from 0.9589 to 0.9509. This indicates that retrieved knowledge does not always help and may introduce noisy or weakly relevant information. With uncertainty-aware fusion, accuracy improves to 0.9529, and the average knowledge weight decreases from 0.3333 to 0.2876. It suggests that uncertainty fusion reduces the negative impact of external knowledge, showing that MUSE is less affected by knowledge-based perturbations. More detailed analysis is provided in Suppl.

5 Limitations

Although the MUSE model demonstrates strong performance and provides transparent explanations for multimodal incorrect information detection, it still has several limitations. First, it depends on the quality of the retrieved knowledge. If the retrieved information unit is noisy, weakly related, or misleading, it may hurt the final prediction. Second, the model compresses each branch into a single-vector representation, which may miss important local text or visual details. Third, the preprocessing and offline feature extraction steps may make the model harder to use in real-world systems. Future work will explore better retrieval filtering, more detailed cross-modal reasoning, and broader evaluation on more platforms and data types.

6 Conclusion

Multimodal incorrect information units substantially disrupt content quality, veracity, and integrity, and it's increasingly crucial to design detection models that extend beyond single-modality reasoning. Compared to existing detection approaches that blend textual content and auxiliary knowledge, we proposed MUSE, a multimodal modular reasoning framework that disentangles content-internal reasoning from knowledge-augmented reasoning. By separating these reasoning pathways, MUSE enables more transparent analysis of how textual, visual, and contextual signals contribute to final predictions. To mitigate noisy knowledge injections, we further introduced an uncertainty-aware fusion strategy to estimate cross-modality uncertainty and calibrate reasoning confidence. Additionally, we defined the VoX metric to quantify the auxiliary knowledge contributions, providing instance-level diagnostics on knowledge impacts. Experimental results show that MUSE achieves competitive performance and delivers interpretable insights on knowledge injections. Importantly, our proposed reasoning framework is backbone-agnostic, providing a flexible foundation for multimodal reasoning.

References

Sara Abdali, Sina Shaham, and Bhaskar Krishnamachari. 2024. Multi-modal misinformation detection: Approaches, challenges and opportunities. *ACM Computing Surveys*, 57(3):1–29.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, and 45 others. 2025. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*.

Bimal Bhattarai, Ole-Christoffer Granmo, and Lei Jiao. 2021. Explainable tsetlin machine framework for fake news detection with credibility score assessment. *arXiv preprint arXiv:2105.09114*.

Christina Boididou, Katerina Andreadou, Symeon Papadopoulos, Duc Tien Dang Nguyen, Giulia Boato, Michael Riegler, Yiannis Kompatsiaris, and 1 others. 2015. Verifying multimedia use at mediaeval 2015. In *Proceedings of CEUR Workshop on MediaEval 2015*, volume 1436.

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*, pages 2897–2905.

Xianfu Cheng, Wei Zhang, Shiwei Zhang, Jian Yang, Xiangyuan Guan, Xianjie Wu, Xiang Li, Ge Zhang, Jiaheng Liu, Yuying Mai, Yutao Zeng, Zhoufutu Wen, Ke Jin, Baorui Wang, Weixiao Zhou, Yunhong Lu, Hangyuan Ji, Tongliang Li, Wenhao Huang, and Zhoujun Li. 2025. Simplevqa: Multimodal factuality evaluation for multimodal large language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4637–4646.

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, pages 7059–7069.

Danilo Croce, Giuseppe Castellucci, and Roberto Basili. 2020. GAN-BERT: Generative adversarial learning for robust text classification with a bunch of labeled examples. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2114–2119.

Wenliang Dai, Junnan Li, DONGXU LI, Anthony Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N Fung, and Steven Hoi. 2023. Instructblip: Towards general-purpose vision-language models with instruction tuning. In *Proceedings of the 37th International Conference on Neural Information Processing System*, pages 49250–49267.

Wenlong Fang, Qiaofeng Wu, Jing Chen, and Yun Xue. 2025. Notes-guided mllm reasoning: Enhancing mllm with knowledge and visual notes for visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19597–19607.

Hiroshi Fukui, Tsubasa Hirakawa, Takayoshi Yamashita, and Hironobu Fujiyoshi. 2019. Attention branch network: Learning of attention mechanism for visual explanation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10697–10706.

Roy Ganz, Yair Kittenplon, Aviad Aberdam, Elad Ben Avraham, Oren Nuriel, Shai Mazor, and Ron Litman. 2024. Question aware vision transformer for multimodal reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13861–13871.

Bin Guo, Yasan Ding, Lina Yao, Yunji Liang, and Zhiwen Yu. 2020. The future of false information detection on social media: New perspectives and trends. *ACM Computing Survery*, 53(4):1–36.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. 2020. Realm: retrieval-augmented language model pre-training. In *Proceedings of the 37th International Conference on Machine Learning*, pages 3929–3938.

Zhenglin Huang, Jinwei Hu, Xiangtai Li, Yiwei He, Xingyu Zhao, Bei Peng, Baoyuan Wu, Xiaowei Huang, and Guangliang Cheng. 2025. SIDA: social media image deepfake detection, localization and explanation with large multimodal model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 28831–28841.

Zhiwei Jin, Juan Cao, Han Guo, Yongdong Zhang, and Jiebo Luo. 2017. Multimodal fusion with recurrent neural networks for rumor detection on microblogs. In *Proceedings of the 25th ACM International Conference on Multimedia*, pages 795–816.

Rohit Kumar Kaliyar, Anurag Goswami, Pratik Narang, and Soumendu Sinha. 2020. Fndnet–a deep convolutional neural network for fake news detection. *Cognitive Systems Research*, 61:32–44.

Dhruv Khattar, Jaipal Singh Goud, Manish Gupta, and Vasudeva Varma. 2019. Mvae: Multimodal variational autoencoder for fake news detection. In *Proceedings of the World Wide Web Conference*, pages 2915–2921.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pages 1746–1751.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, pages 9459–9474.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *Proceedings of the 39th International Conference on Machine Learning*, Proceedings of Machine Learning Research, pages 12888–12900.

Kun Li, George Vosselman, and Michael Ying Yang. 2025. Multimodal rationales for explainable visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 191–201.

Yikang Li, Wanli Ouyang, Xiaogang Wang, and Xiao'ou Tang. 2017. Vip-cnn: Visual phrase guided convolutional neural network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7244–7253.

Junyan Lin, Haoran Chen, Yue Fan, Yingqi Fan, Xin Jin, Hui Su, Jinlan Fu, and Xiaoyu Shen. 2025. Multi-layer visual feature fusion in multimodal llms: Methods, analysis, and best practices. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4156–4166.

Zhiqiu Lin, Samuel Yu, Zhiyi Kuang, Deepak Pathak, and Deva Ramanan. 2023. Multimodality helps unimodality: Cross-modal few-shot learning with multimodal models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19325–19337.

Hu Linmei, Tianchi Yang, Chuan Shi, Houye Ji, and Xiaoli Li. 2019. Heterogeneous graph attention networks for semi-supervised short text classification. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 4821–4830.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, pages 34892–34916.

Giulio Lovisotto, Nicole Finnie, Mauricio Munoz, Chaithanya Kumar Mummadi, and Jan Hendrik Metzen. 2022. Give me your attention: Dot-product attention considered harmful for adversarial patch robustness. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15234–15243.

Jing Ma, Wei Gao, Prasenjit Mitra, Sejeong Kwon, Bernard J. Jansen, Kam-Fai Wong, and Meeyoung Cha. 2016. Detecting rumors from microblogs with recurrent neural networks. In *Proceedings of the 25th International Joint Conference on Artificial Intelligence*, pages 3818–3824.

Qianli Ma, Zhenxi Lin, Jiangyue Yan, Zipeng Chen, and Liuhong Yu. 2020. MODE-LSTM: A parameter-efficient recurrent network with multi-scale for sentence classification. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 6705–6715.

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

Shervin Minaee, Nal Kalchbrenner, Erik Cambria, Narjes Nikzad, Meysam Chenaghlu, and Jianfeng Gao. 2021. Deep learning–based text classification: A comprehensive review. *ACM Computing Survery*, 54(3):1–40.

Mikel K. Ngueajio, Saurav Aryal, Marcellin Atemkeng, Gloria Washington, and Danda Rawat. 2025. Decoding fake news and hate speech: A survey of explainable ai techniques. *ACM Computing Survery*, 57(7):1–37.

Ethan Perez, Florian Strub, Harm de Vries, Vincent Dumoulin, and Aaron Courville. 2018. Film: visual reasoning with a general conditioning layer. In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence and 30th Innovative Applications of Artificial Intelligence Conference and 8th AAAI Symposium on Educational Advances in Artificial Intelligence*, pages 3942–3951.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, and 1 others. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the International Conference on Machine Learning*, pages 8748–8763.

Devendra Singh Sachan, Manzil Zaheer, and Ruslan Salakhutdinov. 2019. Revisiting lstm networks for semi-supervised text classification via mixed objective function. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 6940–6948.

Philipp Schmid and Hannah Bauer. 2025. Impact of exposure to health misinformation on belief in health misinformation: A meta-analysis of rcts. *Health Communication*, 0(0):1–11.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. 2017. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *arXiv preprint arXiv:1701.06538*.

Chongyang Shi, Yijun Yin, Qi Zhang, Liang Xiao, Usman Naseem, Shoujin Wang, and Liang Hu. 2023. Multiview clickbait detection via jointly modeling subjective and objective preference. In *Findings of the Association for Computational Linguistics: EMNLP*, pages 11807–11816.

Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. 2017. Fake news detection on social media: a data mining perspective. *ACM SIGKDD Explorations Newsletter*, 19(1):22–36.

Qi Su, Mingyu Wan, Xiaoqian Liu, and Chu-Ren Huang. 2020. Motivations, methods and metrics of misinformation detection: an nlp perspective. *Natural Language Processing Research*, 1(1):1–13.

Xiaobing Sun and Wei Lu. 2020. Understanding attention for text classification. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3418–3428, Online. Association for Computational Linguistics.

Kjerstin Thorson, Emily Vraga, and Brian Ekdale. 2010. Credibility in context: How uncivil online commentary affects news credibility. *Mass Communication and Society*, 13(3):289–313.

Jens Van Nooten and Walter Daelemans. 2025. Jump to hyperspace: Comparing Euclidean and hyperbolic loss functions for hierarchical multi-label text classification. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 4260–4273.

Jiang Wang, Yi Yang, Junhua Mao, Zhiheng Huang, Chang Huang, and Wei Xu. 2016. Cnn-rnn: A unified framework for multi-label image classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2285–2294.

William Yang Wang. 2017. "liar, liar pants on fire": A new benchmark dataset for fake news detection. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, pages 422–426, Vancouver, Canada. Association for Computational Linguistics.

Yanan Wang, Michihiro Yasunaga, Hongyu Ren, Shinya Wada, and Jure Leskovec. 2023. Vqa-gnn: Reasoning with multimodal knowledge via graph neural networks for visual question answering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 21582–21592.

Yaqing Wang, Fenglong Ma, Zhiwei Jin, Ye Yuan, Guangxu Xun, Kishlay Jha, Lu Su, and Jing Gao. 2018. Eann: Event adversarial neural networks for multi-modal fake news detection. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 849–857.

Yikai Wang, Xinghao Chen, Lele Cao, Wenbing Huang, Fuchun Sun, and Yunhe Wang. 2022. Multimodal token fusion for vision transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12186–12195.

Jiaying Wu, Jiafeng Guo, and Bryan Hooi. 2024. Fake news in sheep's clothing: Robust fake news detection against llm-empowered style attacks. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 3367–3378.

Yang Wu, Pengwei Zhan, Yunjian Zhang, Liming Wang, and Zhen Xu. 2021. Multimodal fusion with co-attention networks for fake news detection. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2560–2569.

Yijin Xiong, Yukun Feng, Hao Wu, Hidetaka Kamigaito, and Manabu Okumura. 2021. Fusing label embedding into BERT: An efficient improvement for text classification. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1743–1750.

Junxiao Xue, Yabo Wang, Yichen Tian, Yafei Li, Lei Shi, and Lin Wei. 2021. Detecting fake news by exploring the consistency of multimodal data. *Information Processing & Management*, 58(5):102610.

Chang Yang, Peng Zhang, Wenbo Qiao, Hui Gao, and Jiaming Zhao. 2023. Rumor detection on social media with crowd intelligence and ChatGPT-assisted networks. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5705–5717.

Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. 2016. Hierarchical attention networks for document classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1480–1489.

Xin Yuan, Zhe Lin, Jason Kuen, Jianming Zhang, Yilin Wang, Michael Maire, Ajinkya Kale, and Baldo Faieta. 2021. Multimodal contrastive training for visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6995–7004.

Joe Yue-Hei Ng, Matthew Hausknecht, Sudheendra Vijayanarasimhan, Oriol Vinyals, Rajat Monga, and George Toderici. 2015. Beyond short snippets: Deep networks for video classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4694–4702.

Huaiwen Zhang, Quan Fang, Shengsheng Qian, and Changsheng Xu. 2019. Multi-modal knowledge-aware event memory network for social media rumor detection. In *Proceedings of the 27th ACM International Conference on Multimedia*, pages 1942–1951.

Zhongxing Zhang, Emily K Vraga, Jisu Huh, and Jaideep Srivastava. 2026. Bimind: A dual-head reasoning model with attention-geometry adapter for incorrect information detection. *arXiv preprint arXiv:2604.06022*.

Yang Zhong and Bhiman Kumar Baghel. 2024. Multimodal understanding of memes with fair explanations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 2007–2017.

Xinyi Zhou, Jindi Wu, and Reza Zafarani. 2020. Similarity-aware multi-modal fake news detection. In *Proceedings of 24th Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining*, pages 354–367.

Xinyi Zhou and Reza Zafarani. 2020. A survey of fake news: Fundamental theories, detection methods, and opportunities. *ACM Computing Survery*, 53(5):1–40.

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

A Methodology Details

Here, more details can be found in the following subsections.

A.1 Feature extraction

For the **in-domain retrieval module**, we perform knowledge retrieval over multimodal representations $\boldsymbol{h}_i^m$ to complement the contextual visual features. Typically, FiLM (Perez et al., 2018) scales (i.e., $1 + \tanh(\boldsymbol{\gamma}_k)$) and shifts (i.e., $\boldsymbol{\beta}_k$) each dimension of textual representations $\boldsymbol{h}_i^c$ and visual signals $\boldsymbol{h}_i^m$, based on the knowledge retrieved from semantic memory. To input visual modality, we design two visual reasoning heads:

$$\mathbf{z}_{\mathrm{M}} = f(\boldsymbol{h}_i^m; \boldsymbol{\theta}_{\mathrm{M}}), \qquad \mathbf{z}_{\mathrm{E}} = f(\boldsymbol{h}_E^m; \boldsymbol{\theta}_{\mathrm{E}}) \quad (16)$$

where $\mathbf{z}_{\mathrm{M}}, \mathbf{z}_{\mathrm{E}} \in \mathbb{R}^K$ are reasoning outputs of the visual-conditioned and knowledge-augmented reasoning heads, respectively. $\boldsymbol{\theta}_{\mathrm{M}}$ and $\boldsymbol{\theta}_{\mathrm{E}}$ are learnable parameters. $\boldsymbol{h}_E^m$ is produced from $\boldsymbol{h}_i^m$ through FiLM layer. Then, we transform $\mathbf{z}_{\mathrm{I}}, \mathbf{z}_{\mathrm{K}}$ into probability distributions through softmax:

$$\hat{\boldsymbol{y}}_{\mathrm{M}} = \mathrm{softmax}(\mathbf{z}_{\mathrm{M}}), \quad \hat{\boldsymbol{y}}_{\mathrm{E}} = \mathrm{softmax}(\mathbf{z}_{\mathrm{E}}) \quad (17)$$

$\hat{\boldsymbol{y}}_{\mathrm{M}}$ is the predictions from the visual-conditioned reasoning $\mathrm{Head}_{\mathrm{M}}$, and it's what the model believes with visual inputs. $\mathrm{Head}_{\mathrm{E}}$ represents knowledge-augmented reasoning and derives $\hat{\boldsymbol{y}}_{\mathrm{E}}$, and $\hat{\boldsymbol{y}}_{\mathrm{E}}$ is what the model believes with external visual knowledge.

Retrieval-augmented VLMs. Beyond the in-domain memory bank retrieval, we deploy retrieval-augmented VLMs that incorporate external multimodal knowledge sources to improve reasoning and factual grounding. Given an information unit (T_i, I_i), retrieval-augmented VLMs first encode the query into a shared embedding space:

$$\boldsymbol{q}_i = \boldsymbol{E}_{\mathrm{VLM}}(T_i, I_i), \quad (18)$$

where $\boldsymbol{E}_{\mathrm{VLM}}(\cdot)$ denotes a pretrained multimodal encoder and $\boldsymbol{q}_i \in \mathbb{R}^{d_s}$ is the query embedding. An external retrieval memory bank is constructed as

$$\mathcal{M} = \{\boldsymbol{m}_1, \boldsymbol{m}_2, \ldots, \boldsymbol{m}_N\}, \quad (19)$$

where $\boldsymbol{m}_j \in \mathbb{R}^{d_s}$ represents a textual, visual, or multimodal embedding from external corpora. The retrieval similarity between the query and memory entries is computed using cosine similarity. Then, the top-k nearest neighbors are selected and aggregated into a retrieval-enhanced representation:

$$\bar{\boldsymbol{m}}_i = \frac{1}{k} \sum_{j \in \mathcal{N}_k(\boldsymbol{q}_i)} \boldsymbol{m}_j. \quad (20)$$

Finally, the retrieved knowledge is fused with the original query representation through retrieval-augmented generation mechanisms:

$$\boldsymbol{h}_i^{\mathrm{aug}} = F(\boldsymbol{q}_i, \bar{\boldsymbol{m}}_i), \quad (21)$$

where $F(\cdot)$ denotes the multimodal fusion function, and the augmented representation $\boldsymbol{h}_i^{\mathrm{aug}}$ is then applied for downstream multimodal reasoning and prediction tasks.

A.2 Prompt Templates

In our implementation, all prompts are used as a single user message and processed through the Qwen-VL chat template before being passed into the frozen multimodal backbone, where we use three prompts for different branches.

Content branch. For content-internal reasoning, we encode the textual content:

```
Is the following news correct or
incorrect? News text: {text}
```

Multimodal branch. For multimodal reasoning, we encode the image together with the news text:

```
Is the following news correct or
incorrect? Use both the news text and
image. News text: {text}
```

This prompt is paired with the visual input and is used to generate the multimodal feature representation.

Visual retrieval branch. For image-based retrieval, we encode the image with the following prompt:

```
Represent this image for visual
retrieval.
```

This prompt is used only to obtain image-retrieval features for querying the visual memory bank, rather than for final classification.

A.3 Uncertainty-aware fusion strategy

In this subsection, we provide foundational details of the proposed uncertainty-aware fusion strategy described in Sec. 3.5 of the main paper. Importantly, we introduce a cross-head refinement mechanism before fusing the knowledge-augmented branch, where the content representation and its

uncertainty signal are extracted to refine the knowledge representation. It helps the MUSE framework to regulate external knowledge using internal content before final fusion. Here, we introduce the Monte-Carlo dropout for uncertainty estimation, the fusion gate construction, and the cross-modal agreement module.

Monte-Carlo dropout for uncertainty estimation. In the main paper, the reasoning uncertainty of each head $h \in \mathcal{H}$ is defined in Eq. (10). To estimate the reasoning uncertainty, we employ Monte-Carlo dropout during inference phase. Specifically, dropout layers remain active, and the MUSE framework performs S stochastic forward passes. For each pass s, the reasoning head produces logits $\mathbf{z}_h^s$, which are converted into class probabilities using the softmax function.

The predictive distribution is then approximated by averaging the probabilities across the stochastic passes:

$$\bar{\mathbf{p}}_h = \frac{1}{S} \sum_{s=1}^{S} \text{softmax}(\mathbf{z}_h^s). \tag{22}$$

The reasoning uncertainty $\mathcal{U}_h$ defined in Eq. (10) is computed as the Shannon entropy of the predictive distribution $\bar{\mathbf{p}}_h$. In our implementation, we set S as 5.

Fusion gate. The fusion weights described in Eq. (12) are produced from a gate feature vector $\mathbf{u}$, which is constructed using uncertainty $\mathcal{U}_h$ and decision margin $\mathcal{M}_h$ from each reasoning head h. The gate vector is then defined as:

$$\mathbf{u} = [\mathcal{U}_{\text{C}}, \mathcal{U}_{\text{M}}, \mathcal{U}_{\text{E}}, \mathcal{M}_{\text{C}}, \mathcal{M}_{\text{M}}, \mathcal{M}_{\text{E}}]. \tag{23}$$

Finally, $\mathbf{u}$ is fed into a two-layer multilayer perceptron (MLP) to generate the fusion weights:

$$\mathbf{w} = \text{softmax}(\text{MLP}(\mathbf{u})), \tag{24}$$

where $\mathbf{w} = \{w_{\text{C}}, w_{\text{M}}, w_{\text{E}}\}$ and $\sum_k w_k = 1$ for each sample x_i.

Fusion of modular reasoning heads. The final fused logits are computed as:

$$\mathbf{z}_{\text{F}} = \sum_{k \in \{\text{C,M,E}\}} \mathbf{w}_k \cdot \mathbf{z}_k \tag{25}$$

where $\mathbf{z}_k$ are logits produced by individual reasoning heads, $\mathbf{w}$ are adaptive fusion weights.

B Experiment Settings

Here, we implement MUSE using PyTorch, where Qwen3-VL-2B model (Bai et al., 2025) serves as the backbone to evaluate the modular and model-agnostic design of the MUSE framework on the Weibo dataset. The detailed hyperparameter settings are reported in Table 4. As shown in Table 4, for both Twitter and Weibo datasets, we set the batch size to 32. Each textual token is represented by a 2048-dimensional embedding vector. For the visual features, each image is represented as a 2048-dimensional vector.

For modular heads, we employ ReLU as an activation function, where modular heads are trained with cross-entropy (CE) loss and a symmetric-KL agreement regularizer. To enhance robustness, we apply modality dropout with a rate of 0.3. Additionally, we adopt class-balanced weights on the CE loss to handle the class imbalance issue. The final output layer with a softmax function is designed to provide the probability distribution indicating the likelihood of the content being labeled 1 (i.e., correct information) or 0 (i.e., incorrect information). In the uncertainty estimation, the key hyperparameter settings are set as follows: $S = 5$, dropout rate is 0.3, gate MLP hidden dimension is 64, and agreement weight $\lambda_{agree} = 0.05$. Here, we use Adam optimizer 1×10^{-4} to train our model, where we employ early-stopping (patience = 7) to tune the hyperparameters.

In our training setup, the pretrained Qwen backbone is frozen, and only the classifier parameters are updated. The full pipeline has approximately 1.25B parameters, including 1.22B frozen backbone parameters and 32.72M trainable classifier parameters. Thus, only about 2.61% of the total parameters are trainable.

For training and testing our MUSE and other approaches, we follow the original dataset setup and split the training portion into 80% for training and 20% for validation. All the experiments were conducted on 1 NVIDIA A100 GPU with 40 GB RAM. For the Evaluation Metrics, we quantitatively evaluated our model's performance compared to the other 14 tested models, using classification metrics such as accuracy (Acc), Macro-precision (Pre), Macro-F1 (F1), and Macro-recall (Rec). Table 1 reports the main benchmark comparison; other tables (e.g., Tables 2 and 3) are designed to analyze relative component impact within these tables under different experimental environments.

Training		Model Architecture	
Hyperparameter	**Value**	**Hyperparameter**	**Value**
Batch size	32	Base model	Qwen3-VL-2B-Instruct
Learning rate	1×10^{-4}	Qwen hidden dim	2048
Weight decay	1×10^{-4}	ReasoningHead hidden dim	512
Optimizer	Adam	ReasoningHead intermediate dim	128
Auxiliary head weight	0.3	Number of classes	2
Number of epochs	15	Dropout rate	0.3
Early stopping patience	7	Quantization	4-bit (NF4)
Train/val split	80/20	Number of reasoning heads	3
Random seed	1	Backbone precision	float16
Uncertainty / Fusion		**Feature Extraction**	
Hyperparameter	**Value**	**Hyperparameter**	**Value**
MC dropout passes	5	Content branch	Text-only prompt
λ_{agree}	0.05	Multimodal branch	Text+Image prompt
λ_{VoX}	0.01	Knowledge branch	Text+Image prompt
Fusion hidden dim	64	Feature token	Last hidden state
VoX regularization	Off	Frozen backbone	Yes

Table 4: Hyperparameter settings for the Qwen3-VL-based MUSE model on the Weibo dataset.

C Experiment Results

C.1 Quantitative results

For further evaluating the cross-modality reasoning capability of our MUSE framework, we compared the Qwen-based MUSE model with the standard CLIP model (Radford et al., 2021) and several vision-language models (VLMs), including BLIP (Li et al., 2022), InstrcutBLIP (Dai et al., 2023), LLaVA-1.5-7B (Liu et al., 2023), and Qwen3-VL-7B (Bai et al., 2025). The experimental results in Table 5 demonstrate that MUSE with Qwen3-VL-2B backbone achieves the best performance among all evaluated multimodal models, reaching an accuracy of 0.962, which surpasses strong multimodal VLM baselines such as Qwen3-VL-7B (0.954) and LLaVA-1.5-7B (0.951). In addition to higher accuracy, MUSE has the highest F1-score for incorrect information detection (0.963) across both classes, outperforming other tested models. Compared with Qwen3-VL-7B, MUSE significantly improves precision for detecting incorrect information (0.957 vs. 0.931) and reduces false positive predictions by approximately 40%, indicating that MUSE is less likely to incorrectly label correct content as incorrect information. Additionally, MUSE has a higher recall for incorrect information (0.969) over the Qwen3-VL-7B model. We can see that CLIP has the highest precision (1.000) but much lower recall (0.834), showing that it misses many incorrect samples due to its similarity-based decision mechanism. BLIP (Pre 0.942, Rec 0.883) enhances cross-modal interaction, but it still lacks the strong reasoning capability. In addition, InstructBLIP (Pre 0.909, Rec 0.934) improves recall through instruction tuning but introduces more false positives. Although LLaVA, based on a multimodal LLM, receives a more balanced performance (Pre 0.953, Rec 0.949), it falls short of MUSE in both overall accuracy and incorrect information F1.

To isolate the architecture-level gains from backbone effects, we reimplement CAFE model with the same frozen Qwen3-VL-2B model; this unified-backbone CAFE receives 92.6% accuracy and 92.7% F1 on Weibo, improving the original baseline in Table 1. However, MUSE achieves 96.2% accuracy and 96.3% F1, suggesting that the remaining gains arise from the proposed modular reasoning design. In addition, the uncertainty-aware fusion provides additional robustness under noisy knowledge retrieval, as shown in the ablation study.

The Qwen3-VL baseline performs poorly in both prompting settings, with zero-shot reaching 52.15% accuracy and few-shot improving to 63.08% accuracy. It has a bias toward predicting the incorrect class, suggesting that raw VLM generation is not reliable for this task. Though the few-shot setting reduces the prediction imbalance, it's still far below the trained MUSE variants.

These results highlight the merits of the MUSE modular reasoning architecture over multimodal VLMs. Although VLMs rely on a unified representation to produce predictions, MUSE decomposes reasoning into content, visual, and knowledge reasoning, followed by an uncertainty-aware

Method	Acc	Incorrect Information Unit			Correct Information Unit		
		Pre	Rec	F1	Pre	Rec	F1
CLIP (Radford et al., 2021)	0.917	**1.000**	0.834	0.910	0.857	**1.000**	0.923
BLIP (Li et al., 2022)	0.914	0.942	0.883	0.912	0.890	0.946	0.917
InstructBLIP (Dai et al., 2023)	0.920	0.909	0.934	0.921	0.932	0.906	0.919
LLaVA-1.5-7B (Liu et al., 2023)	0.951	0.953	0.949	0.951	0.949	0.953	0.951
Qwen3-VL-7B (Bai et al., 2025)	0.954	0.931	**0.981**	0.955	**0.980**	0.927	0.953
Qwen3-VL-7B$_{\text{zero-shot}}$ (Bai et al., 2025)	0.522	0.512	0.988	0.674	0.815	0.053	0.100
Qwen3-VL-7B$_{\text{few-shot}}$ (Bai et al., 2025)	0.631	0.582	0.936	0.718	0.835	0.324	0.467
CAFE-Qwen (Chen et al., 2022)	0.926	0.917	0.938	0.927	0.936	0.915	0.925
MUSE	**0.962**	0.957	0.969	**0.963**	0.969	0.956	**0.962**

Table 5: Performance comparison on the Weibo dataset. The best results are highlighted in **bold**.

fusion mechanism that calibrates the final predictions. This structured reasoning design allows the model to balance complementary signals and avoid contrastive predictions, providing more balanced class performance and better decision calibration and reducing both false positives and false negatives. In summary, these results demonstrate that structured modular reasoning and calibrated fusion strategy can outperform larger multimodal VLMs on the incorrect information detection task, achieving higher precision, better-balanced predictions, and improved interpretability.

C.2 Ablation study

To validate the design choices in MUSE, we compared our retrieval and fusion mechanisms with several controlled baselines, as shown in Table 6. For retrieval, we evaluated a RAG-style concatenation baseline and a kNN classifier without feature modulation. Although naive RAG concatenation achieves comparable accuracy (96.34%), the kNN model suffers a significant drop in performance (95.24%), suggesting that retrieval alone is insufficient. MUSE has strong performance and reduces false-positive predictions compared with RAG-Concat, with a 6.5% reduction and improving over the kNN baseline, which produces more false positives (37.7% higher than MUSE). These results suggest that retrieval-conditioned reasoning has more stable and reliable predictions than naive retrieval augmentation.

For fusion, we compared our uncertainty-aware strategy with a standard Mixture-of-Experts (MoE) model (Shazeer et al., 2017) using learned gating. Although the MoE baseline achieves competitive accuracy (96.24%), it presents more false-positive errors, which shows that the expert routing tends to learn a dominant expert and overtrust individual reasoning head. In contrast, our uncertainty-aware fusion reduces 24.6% false positives compared with MoE and balances predictions across reasoning heads, providing better calibration and more reliable decisions. In summary, MUSE introduces interpretable uncertainty signals and modality-specific reasoning heads, enabling better diagnostic analysis of model decisions. These properties make MUSE more scalable and interpretable than retrieval-based baselines.

C.3 VoX analysis

To verify that VoX reflects knowledge contributions rather than a trivial logit difference, we conducted 3 validation experiments as shown in Table 7. First, we performed a faithfulness test by removing retrieved knowledge at inference time. Samples with high VoX have a significantly larger accuracy drop (-5.81%) than low-VoX samples (-1.00%), suggesting that VoX correctly identifies cases where auxiliary knowledge is important. A permutation test ($p < 0.001$) further verifies that this observation is statistically significant. Second, we conducted a counterfactual retrieval experiment by replacing retrieved neighbors with randomly permuted ones. It degrades accuracy from 96.5% to 70.3% and flips the mean VoX from +0.52 to -0.87, indicating that VoX tracks whether retrieved experience contributes positively or negatively to the final predictions. Third, we analyzed cases where the content-internal head is incorrect, but the fused prediction becomes correct. Among these corrections, 78.9% are related to positive VoX signals, showing that VoX identifies cases where external knowledge improves predictions. In summary, VoX provides a logit-level diagnostic signal, quantifying how auxiliary knowledge shifts the model's prediction.

Retrieval noise robustness. To further evaluate whether the knowledge-augmented reasoning head

Category	Method	Acc	Prec	Rec	F1
Retrieval	RAG-Concat (retrieve+concat+MLP)	0.9634	0.9549	0.9730	0.9638
	kNN (retrieved-only, no FiLM)	0.9524	0.9338	0.9740	0.9535
Fusion	Standard MoE (learned gating)	0.9624	0.9451	0.9820	0.9632
	MUSE (Uncertainty-aware fusion)	**0.9634**	**0.9576**	0.9700	0.9637

Table 6: Ablation study on retrieval strategies and fusion mechanisms.

Experiment	Metric	Result
Faithfulness Test	Acc Drop (High-VoX)	5.81%
	Acc Drop (Low-VoX)	-1.00%
	Permutation Test	$p < 0.001$
Counterfactual Retrieval	Accuracy Change	$0.965 \rightarrow 0.703$
	Mean VoX Change	$0.522 \rightarrow -0.867$
	Confidence Change	$0.957 \rightarrow 0.884$
Correction Analysis	VoX-positive Samples	1183 / 1996
	Correction Rate	14.2%
	Precision of VoX Signal	78.9%

Table 7: Empirical validation of VoX, including faithfulness test, counterfactual retrieval, and correction analysis.

Noise p	Acc	F1	Mean VoX
0.00	**0.9654**	**0.9654**	+0.522
0.25	0.8947	0.8945	+0.170
0.50	0.8320	0.8312	-0.178
0.75	0.7663	0.7642	-0.501
1.00	0.7036	0.6989	-0.851

Table 8: Robustness under retrieval corruption. A proportion p of retrieved neighbors is replaced with random posts.

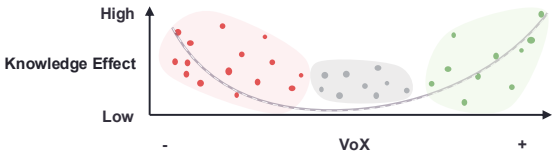

Figure 3: An illustration of knowledge impacts at the instance level.

benefits from semantical retrieval, we progressively corrupt the retrieved neighbor set by replacing a proportion p of retrieved samples with randomly selected posts from the corpus. As shown in Table 8, detection performance degrades as retrieval noise increases, with accuracy dropping from 0.9654 under normal retrieval to 0.7036 when all neighbors are randomized. It shows that the knowledge branch relies on meaningful retrieval signals rather than simple feature augmentation. In addition, the VoX score tracks retrieval quality, decreasing from +0.522 to -0.851, indicating that VoX can reflect whether external knowledge improves or harms the final predictions.

VoX visualization. To further interpret these VoX values, we visualized 3 types of knowledge effects in Figure 3, illustrating how auxiliary knowledge can improve predictions, remain neutral, or hurt performance (as in negative VoX-gain routes) across samples. Negative VoX value suggests that knowledge may introduce noisy or harmful shifts, while positive VoX indicates that knowledge helps improve predictions. Near-zero VoX denotes neutral cases, where auxiliary knowledge has limited effects. In this figure, the y-axis represents knowledge effect magnitude, where larger negative or positive VoX values have stronger prediction changes, forming a U-shaped pattern, whereas neutral knowledge effects remain close to the bottom.

In addition, the VoX correlation analysis shows that VoX exhibits strong positive correlations with accuracy across modular heads, with the final fused head ($\mathbf{z}_{\mathrm{F}}$) achieving the highest Pearson (≈ 0.7) and Spearman correlations. It indicates that, compared to the content-internal reasoning head ($\mathbf{z}_{\mathrm{C}}$), different heads capture complementary features: visual reasoning ($\mathbf{z}_{\mathrm{M}}$) and knowledge-augmented reasoning ($\mathbf{z}_{\mathrm{E}}$). In addition, from the Pearson and Spearman correlations of $\mathbf{z}_{\mathrm{F}}$, we can see that our proposed uncertainty-aware cross-modal strategy help reasoning heads handle cross-modality inconsistency and balance detection performance, en-

hancing model robustness. The strong correlations between VoX and accuracy demonstrate modality complementarity, where the VoX-correctness relationship proves that our uncertainty-aware strategy can identify when multi-modal reasoning helps prediction. These findings suggest that our MUSE's core novelty is integrating cross-modality features and capturing predication uncertainty to improve correctness-oriented reasoning.

C.4 Artifact licenses

In this work, we only use the existing artifacts for academic research and model evaluation purposes. The datasets, such as Twitter and Weibo, are used as benchmark datasets. For the pretrained backbone models (e.g., Qwen3-VL-2B) and comparison model artifacts (e.g., EANN) are used under their original licenses or terms of use. We do not redistribute third-party datasets, pretrained models, or baseline code. All these artifacts used in this work can be accessed from their original public sources, following their licenses, terms of use, and distribution requirements.

C.5 Data leakage

We perform several safeguards to avoid data leakage issues. Specifically, the training and test sets are loaded from separate files, validation data are split from the training set, and internal retrieval memory is constructed from the training subset rather than from validation or test samples. For in-domain retrieval modes, it masks the self-matches using sample identifiers, with separate ID ranges assigned to training, test, and external knowledge samples to avoid leakage.

Knowledge Safeguards. In addition, MUSE employs knowledge safeguards across retrieval, training, and fusion to reduce the impact of noisy or unreliable retrieved knowledge. We summarize these safeguards as follows:

- Similarity-aware retrieval: We first compute top-k retrieval similarity statistics and derive a retrieval uncertainty signal, where this signal is used by the fusion module to reduce the impact of noisy knowledge.

- Bounded FiLM injection: Retrieved knowledge is then injected through FiLM modulation, where the scaling term is bounded by $tanh$ to reduce feature distortion from noisy retrieval.

- Agreement loss: A pairwise agreement regularizer is designed to penalize divergence among the active reasoning heads, helping stabilize training when retrieved knowledge is noisy or inconsistent.

- Retrieval isolation safeguards: For in-domain retrieval modes, the memory bank is built from the training subset, and the self-retrieval is blocked using sample identifiers.