Exploring Hyperbolic Hierarchical Structure for Multimodal Rumor Detection

Md Mahbubur Rahman [1]*, Shufeng Hao[2] * , Chongyang Shi[1] † , An Lao[1], Jinyan Liu[1] †
[1]School of Computer Science and Technology, Beijing Institute of Technology
[2]College of Computer Science and Technology (College of Data Science),
Taiyuan University of Technology
{rahmanmahbub073, cy_shi, an.lao, jyliu}@bit.edu.cn, haoshufeng@tyut.edu.cn

Abstract

The rise of multimodal content on social platforms has led to the rapid spread of complex and persuasive false narratives, combining of text and images. Traditional rumor detection models attempt to identify such content by relying on textual cues or employing shallow multimodal fusion techniques. However, these methods often assume a simplistic one-to-one alignment between modalities, overlooking the richer hierarchical relationships across modalities, failing to capture the layered structure of meaning. In this paper, we present RumorCone, a novel method that employs hyperbolic geometry in order to preserve hierarchical, non-linear relationships, rather than representing them at a flat semantic level. First, RumorCone decomposes image and text content into three levels: base, mid, and high-level abstractions, and embeds them in hyperbolic space to model their tree-like semantic structure. Second, a dynamic hyperbolic multimodal attention mechanism aligns features across modalities and levels, and a flexible fusion strategy adjusts the contribution of each modality based on alignment quality. Our experiments indicate the importance of hierarchical semantic modeling for robust and interpretable multimodal rumor detection.

1 Introduction

The rapid spread of rumors across social media platforms has emerged as a serious societal challenge, particularly due to the complex nature of multimodal content that integrates both textual and visual elements. Recent advancements in multimodal rumor detection have demonstrated potential in utilizing multiple data sources; yet, current models often fail to sufficiently represent the semantic interactions between modalities, particularly in aligning images with corresponding texts. However, a critical limitation is their failure to model the hierarchical structure of visual semantics, which is essential for correctly interpreting an images meaning in conjunction with its textual context.

Previous approaches, such as HSEN (Zhang et al., 2023), HAGNN (Xu et al., 2023b), and HMCAN (Qian et al., 2021) have introduced hierarchical learning mechanisms to enhance image-text alignment. These models look for to establish connections between textual and visual modalities; they frequently model low-level interdependencies by treating the visual-linguistic paradigms as one-to-one correspondences. In contrast, this method neglects the rich dependencies that emerge when the two modalities are looked at from a variety of hierarchical semantic viewpoints. More specifically, these models often ignore the complex, non-linear, tree-like relationships that frequently define real-world information in favour of treating semantic features in a flat or linear manner. This restriction is especially problematic in posts that are emotionally charged or crisis-related, as meaning is conveyed through layers of abstraction that range from fine-grained factual details to general emotional cues. Flattened representations collapse these distinctions, leading to semantic mismatches and reducing model sensitivity to subtle misinformation cues.

This challenge is clearly illustrated by the example in Fig. 1, which demonstrates the importance of tree-like semantic hierarchies. In the Non-rumor post, which depicts a real earthquake in Morocco, both the image and text features exhibit strong consistency across three levels of abstraction. The image hierarchy progresses from low-level cues (e.g., rubble, dust) to mid-level objects (e.g., collapsed buildings) and finally to high-level semantic context (e.g., natural disaster). The textual hierarchy mirrors this: from base-level emotional cues (e.g., shocking or tragic), to mid-level

*Equal contribution.
†Corresponding author.

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 115–134
November 4-9, 2025

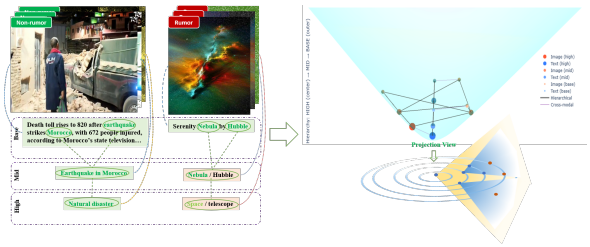

Figure 1: Image and text features are modeled as tree-like structures across three levels: base (low-level cues), mid (object meanings), and high (overall context). In the non-rumor, image and text are well aligned across all levels. In the rumor, they are misaligned, especially at higher levels. Hyperbolic space captures and preserves these relationships inherent in multimodal content, enabling the detection of semantic misalignments that are characteristic of rumors.

descriptions of the event (e.g., earthquake in Morocco), and finally to abstract or factual framing (e.g., references to disaster response or regional impact). In contrast, the Rumor post pairs a visually unrelated image of a nebula with text about a space-related event. While the base-level visual (e.g., colors and light) may loosely attract attention, its mid- and high-level semantics do not align with the textual claims, which themselves may escalate from generic surprise to fabricated factual claims. These ambiguities, across both image and text hierarchies, highlight a core issue in multimodal rumor detection: as the level of abstraction increases, the semantic alignment between image and text either strengthens or breaks down, significantly impacting model performance.

This observation motivates the following hypothesis:

> **H1:**Semantic hierarchies between modalities are more likely to align or misalign as abstraction levels increase, leading to more accurate classification of rumors and nonrumors.

We argue that effectively detecting such nuanced manipulations requires explicit modeling of semantic hierarchies, structured relationships that connect an image to textual descriptions across varying abstraction layers. Unlike existing approaches that focus on shallow feature fusion, this perspective emphasizes the importance of semantic structure in distinguishing authentic from deceptive content.

To address these challenges, we propose **RumorCone**, a novel multimodal rumor detection framework that takes advantage of hierarchical semantic alignment and hyperbolic geometry. Our approach makes three primary contributions:

- We introduce a hierarchical semantic alignment module that explicitly models the image-caption relationship across three levels: generic emotional impression, mid-level description, and specific factual detail. This multi-layered perspective allows us to track semantic consistency by preserving non-linear, tree-like relationships in line with our hypothesis (**H1**).
- We propose a novel cross-modal fusion strategy in hyperbolic space, where text and image features are combined at different levels of abstraction. By using attention mechanisms in hyperbolic space, our model prioritizes the most relevant modality for each rumor detection task, resulting in better performance than traditional shallow feature fusion techniques.
- We employ hyperbolic geometry, incorporating three specialized pathways: unimodal processing, cross-modal attention interaction, and hierarchical semantic correlation, preserving their hierarchical relationships. Additionally, we design an adaptive feature modulation mechanism that adjusts the contribution of each modality based on its relevance, improving the model's ability to detect subtle patterns.

2 Method

Our RumorCone[1] framework addresses our hypothesis by explicitly modeling the hierarchical relationships between image and text content across multiple levels of semantic abstraction. Unlike prior work that relies on flat semantic representations, RumorCone leverages hyperbolic geometry to preserve the natural tree-like structure of semantic relationships. The proposed architecture, shown in Fig. 2 , consists of three main components: (1) multimodal feature extraction at multiple abstraction levels, (2) Multi-Modal Feature Fusion in Hyperbolic Space (3) Parallel Processing Pathways for rumor detection.

[1]The name RumorCone is inspired by MERUs entailment learning method (Desai et al., 2023), where image features are guided to stay within a cone-shaped region based on text features.

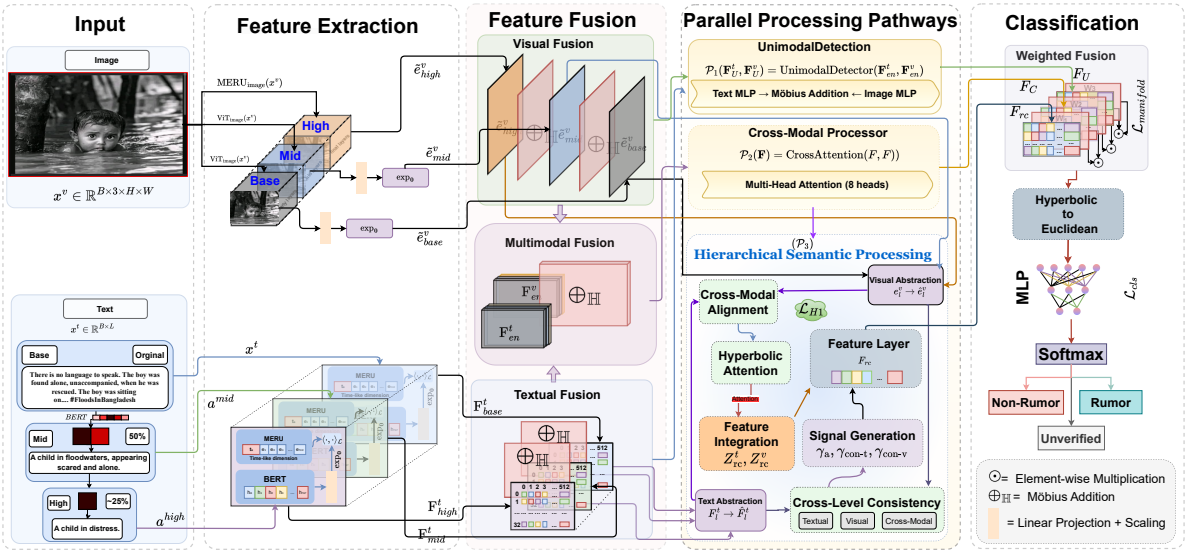

Figure 2: Overview of the RumorCone framework for multimodal rumor detection. The framework processes text and image content through multiple semantic abstraction levels in hyperbolic space, measuring cross-modal alignment and consistency to detect rumors. The hyperbolic embedding approach preserves the hierarchical nature of semantic relationships, enabling detection of inconsistencies across abstraction levels that characterize manipulated content.

2.1 Problem Definition

Let $\mathcal{D} = \{(x_i^t, x_i^v, y_i)\}_{i=1}^N$ denote a dataset of N social media posts, where x_i^t represents the textual content, x_i^v represents the visual content (image), and $y_i \in \{0, 1, 2\}$ is the class label with 0 for rumors, 1 for non-rumors, and 2 for unverified content. Addressing the limitations of flat semantic representations identified in our introduction, we incorporate multiple semantic abstraction levels for each post, denoted as $\{a_i^{mid}, a_i^{high}\}$, representing concrete, mid-level, and high-level semantic abstractions of the content. The objective is to learn a function $f : (X^t, X^v, A) \rightarrow Y$ that maps the multimodal inputs and their abstractions to the correct rumor classification.

2.2 Multi-Modal Feature Extraction

Hierarchical Textual Feature Extraction. To enable hierarchical semantic reasoning in text, we construct three abstraction levels: base (original), mid-level (50% compressed), and high-level (25% compressed). Then, we employ the pre-train language model BERT and an multi-layer perceptron to obtain the correpsonding Euclidean embeddings h_l^t for each abstraction $l \in (base, mid, high)$. Then we employ Lorentzian exponential mapping $\exp_{\mathbf{0}}$ as used in MERU (Desai et al., 2023) to obtain text representation in Lorentzian hyperbolic space, defined as:

$$h_{\text{hyp-}l}^t = \exp_{\mathbf{0}}(h_l^t \cdot \alpha_{\text{exp}}), \tag{1}$$

Meanwhile, we employ a MERU-style transformer encoder to extract complementary textual features e_l^t for each abstraction $l \in (base, mid, high)$. And we also employ Lorentzian exponential mapping to obtain the MERU-style textual representation $e_{\text{hyp-}l}^t$. These embeddings are naturally suited for hyperbolic geometry and further enrich hierarchical text representations.

Hierarchical Image Feature Extraction. For image content x^v, we use early and middle layers of a shared Vision Transformer (ViT) to obtain low-level e_{base}^v and mid-level features e_{mid}^v, capturing local and intermediate visual semantics. To ensure a unified hyperbolic geometry across all abstraction levels, we explicitly project and map them into hyperbolic space using Lorentzian exponential mapping:

$$\tilde{e}_l^v = \exp_0(\alpha_l^v \cdot W_l^v \cdot e_l^v) \tag{2}$$

where, $\exp_0$ is the Lorentzian exponential map to hyperbolic space, α_l^v is a learnable scaling parameter, W_l^v is the projection matrix for level l. And we employ the MERU model to directly obtain the high-level hyperbolic embeddings $\tilde{e}_{\text{high}}^v$, which

uses a ViT backbone and outputs high-level representations already embedded in Lorentzian hyperbolic space. Thus, this hierarchical representation in hyperbolic space allows our model to capture both local and global visual semantics while preserving their hierarchical relationships through the geometry of the space.

2.3 Multi-Modal Feature Fusion in Hyperbolic Space

Intra-Modal Fusion in Hyperbolic Space. For each modality, we aggregate hyperbolic embedding representation across three levels by using hyperbolic Möbius addition operation $\oplus_{\mathbb{H}}$ in the hyperbolic space, defined as:

$$F_{en}^{m} = F_{base}^{m} \oplus_{\mathbb{H}} F_{mid}^{m} \oplus_{\mathbb{H}} F_{high}^{m} \tag{3}$$

where $m \in \{t, v\}$. Here we fuse the textual representation h_{hyp-l}^{t} and e_{hyp-l}^{t} derived from BERT and MERU to obtain textual hyperbolic embedding $F_{l}^{t} = \langle h_{hyp-l}^{t}, e_{hyp-l}^{t} \rangle_{\mathbb{L}}$, where $\langle , \rangle_{\mathbb{L}}$ denotes the Lorentzian inner product. The visual hyperbolic embedding F_{l}^{v} for each abstract level is the hierarchical image feature $\tilde{e}_{l}^{v}$.

Cross-Modal Fusion in Hyperbolic Space. To integrate the hierarchical semantic features from both modalities, we employ an attention-based cross-modal fusion mechanism to obtain the fused representation F_{fused}, computed as:

$$F_{\text{fused}} = \sigma(\alpha_1) \cdot f_{\text{attn}}(F_{\text{en}}^{t}) + \sigma(\alpha_2) \cdot f_{\text{attn}}(F_{\text{en}}^{v}), \tag{4}$$

where, $\sigma(\alpha_1)$ and $\sigma(\alpha_2)$ are learnable parameters that control the attention weights for the text and image features, and the attention function $f_{\text{attn}}(z) = \text{GELU}(\text{LN}(W_{\text{attn}} z + b_{\text{attn}}))$.

2.4 Parallel Processing Pathways

2.4.1 Unimodal Detection Processing

We employ modality-specific processing pathways to extract discriminative features from each input stream independently before cross-modal fusion. The unimodal detector module processes extract discriminative features from each modality, which is formulated as $F_{U}^{m} = \text{MLP}(F_{en}^{m})$, where $m \in \{a, t\}$, and MLP denotes a two-layer neural network with batch normalization and ReLU activations. Then we aggregated the features from individual modalities to obtain the fused unimodal representation $F_U = F_{U}^{t} \oplus_{\mathbb{H}} F_{U}^{v}$.

This approach preserves modality-specific information that might otherwise be lost during cross-modal fusion, contributing to the model's robustness when one modality carries more discriminative information than the other. This helps capture complementary unimodal patterns that are critical when one modality (e.g., text) is unreliable.

2.4.2 Cross-Modal Pathway Processing

The Cross Modal Processor (CMP) module employs hyperbolic multi-head attention to directly model modality-aware interaction in Lorentzian space. The fused embedding is processed with gyrovector operations, defined as:

$$\text{F}_C = F_{\text{fused}} \oplus_{\mathbb{H}} \text{Dropout}(\text{MHA}(F, F, F)) \tag{5}$$

where $F = \text{LN}(F_{\text{fused}})$, and LN() denotes normalization while preserving hyperbolic geometry. $\text{MHA}(Q, K, V) = \text{Concat}(\text{h}_1, \ldots, \text{h}_h)$ is the hyperbolic multi-head attention function, where $\text{h}_i = \text{Attention}(Q \otimes_{\mathbb{H}} W_i^{Q}, K \otimes_{\mathbb{H}} W_i^{K}, V \otimes_{\mathbb{H}} W_i^{V})$, $\otimes_{\mathbb{H}}$ denotes the Möbius matrix multiplication in hyperbolic space and h is the number of attention heads (set to 8 in our implementation).

In the context of our hypothesis H1, the Cross Modal Processor helps identify misalignments between textual and visual modalities across different levels of abstraction by enabling rich cross-modal interactions in the representation space.

2.4.3 Hierarchical Semantic Processing

The hierarchical semantic processing mechanism processes both text and image features at multiple levels of abstraction (Base, Mid, High) to detect the characteristic inconsistency patterns in rumors. The framwork of hierarchical semantic processing is shown in Figure 3.

Hyperbolic Hierarchical Abstraction. These representations at each level (F_{l}^{t}, e_{l}^{v}) are derived from dedicated hyperbolic projection heads, as defined in Sections 2.2 and 2.3. This structure allows RumorCone to perform level-specific alignment using Lorentzian metrics, ensuring full compatibility with the hierarchical modeling required by Hypothesis H1. The abstraction representation for each modality is computed as:

$$\hat{F}_{l}^{m} = \text{GELU}(\text{LN}(W_l h + b_l)) \tag{6}$$

where $l \in \{\text{base}, \text{mid}, \text{high}\}$ denotes each abstraction level and $m \in \{t, v\}$. These abstraction levels directly model the hierarchical nature of semantics, allowing us to represent content from generic emotional impressions to specific factual details.

Cross-Modal Alignment. Following our hypothesis (H1), the cross-modal alignment module employs a cross-modal hyperbolic attention mechanism to dynamically compute how text and image features should attend to each other at each level. The aligned features between text and image at each abstraction level are computed as: $A_{\text{t},l} = \text{HypAtt}(\hat{F}_l^t, \hat{e}_l^v)$ and $A_{\text{v},l} = \text{HypAtt}(\hat{e}_l^v, \hat{F}_l^t)$. The hyperbolic attention mechanism is defined as:

$$\text{HypAtt}(q,k) = \sum_j \frac{\exp(-d_{\mathbb{H}}(q,k_j)/\tau)}{\sum_{j'}\exp(-d_{\mathbb{H}}(q,k_{j'})/\tau)} \otimes_{\mathbb{H}} k_j \tag{7}$$

where $d_{\mathbb{H}}$ represents hyperbolic distance and $\otimes_{\mathbb{H}}$ represents Möbius scalar multiplication.

Cross-Level Consistency. Cross-level consistency ensures that textual and image features remain aligned across multiple abstraction levels. The module computes the hierarchical consistency by using the hyperbolic distance between text and image features across different levels. First, the uni-modality consistency across abstraction levels is computed as : $s_{\text{base,mid}}^{\text{consist-m}} = d_{\mathbb{H}}(\hat{F}_{\text{base}}^m, \hat{F}_{\text{mid}}^m)$ and $s_{\text{mid,high}}^{\text{consist-m}} = d_{\mathbb{H}}(\hat{F}_{\text{mid}}^m, \hat{F}_{\text{high}}^m)$, where $m \in \{a,t\}$. Second, the cross-modality consistency across abstraction levels is calculated as : $s_l = d_{\mathbb{H}}([\hat{F}_l^t; \hat{e}_l^v])$, where $l \in \{base, mid, high\}$.

This ensures that textual and visual features remain consistent across levels, which is critical for detecting rumors that tend to show inconsistencies between features across different levels. This is because that inconsistencies across these levels often indicate manipulated or fabricated content.

Bi-directional Hyperbolic Feature Integration. To fully utilize both directions of cross-modal attention mechanisms, we integrate not only the abstracted features but also the attention outputs from cross-modal alignment. This creates a more balanced representation that captures both text-to-image and image-to-text alignment patterns:

$$\begin{aligned} Z_{\text{rc}}^t &= \bigoplus_{\mathbb{H},l} \alpha_l \otimes_{\mathbb{H}} \hat{F}_l^t \bigoplus_{\mathbb{H},l} A_{\text{t},l}, \\ Z_{\text{rc}}^v &= \bigoplus_{\mathbb{H},l} \beta_l \otimes_{\mathbb{H}} \hat{e}_l^v \bigoplus_{\mathbb{H},l} A_{\text{v},l}, \end{aligned} \tag{8}$$

where $l \in \{\text{base}, \text{mid}, \text{high}\}$, $\bigoplus_{\mathbb{H}}$ represents iterative Möbius addition, $\otimes_{\mathbb{H}}$ is Möbius scalar multiplication, and the attention weights are calculated as:

$$\begin{aligned} \alpha_l &= \frac{\exp(-d_{\mathbb{H}}(c_t, \hat{F}_l^t)/\tau)}{\sum_{l'}\exp(-d_{\mathbb{H}}(c_t, \hat{F}_{l'}^t)/\tau)}; \\ \beta_l &= \frac{\exp(-d_{\mathbb{H}}(c_v, \hat{e}_l^v)/\tau)}{\sum_{l'}\exp(-d_{\mathbb{H}}(c_v, \hat{e}_{l'}^v)/\tau)}, \end{aligned} \tag{9}$$

where c_t and c_v are learnable hyperbolic centroids and α_l and β_l are learned attention weights.

This enhancement ensures that both directions of cross-modal attention contribute to the final representation. The textual features Z_{rc}^t now incorporate not only weighted abstracted text features but also attention-weighted visual contexts that are relevant to each textual component. Similarly, visual features Z_{rc}^v incorporate both weighted abstracted visual features and attention-weighted textual contexts relevant to each visual component.

Consistency and Alignment Signal Processing. Once the alignment and consistency between textual and visual features are obtained, we generate semantic signals. The alignment and consistency scores from H1 are fed into the Semantic Signal Network. We generate modulation factors from the consistency and alignment signals:

$$\gamma_{\text{a}}, \gamma_{\text{t}}, \gamma_{\text{v}} = \sigma(\text{HypMLP}(s_{\text{a}}, s_{\text{t}}, s_{\text{v}}) \tag{10}$$

where σ is Sigmoid activation function, $s_{\text{a}} = [s_{\text{base}}, s_{\text{mid}}, s_{\text{high}}]$, $s_{\text{t}} = [s_{\text{base,mid}}^{\text{consist-t}}, s_{\text{mid,high}}^{\text{consist-t}}]$, $s_{\text{v}} = [s_{\text{base,mid}}^{\text{consist-v}}, s_{\text{mid,high}}^{\text{consist-v}}]$. HypMLP operates directly in the hyperbolic manifold, defined as:

$$\begin{aligned} \text{HypMLP}(x) &= \sigma_{\mathbb{H}}(W \otimes_{\mathbb{H}} x \oplus_{\mathbb{H}} b), \\ \sigma_{\mathbb{H}}(x) &= \exp_0(\sigma(\log_0(x))), \end{aligned} \tag{11}$$

where $\log_0$ represents the logarithmic map from the hyperbolic space to the Euclidean space.

Semantic Feature Layer. The semantic feature layer module combines multiple feature transformations to obtain a unified representation. The formulation is defined as:

$$\begin{aligned} F_{\text{rc}} = \beta_1 \cdot \gamma_{\text{a}} \odot \text{HypMLP}(Z_{\text{rc}}^t \oplus_{\mathbb{H}} Z_{\text{rc}}^v) + \\ \beta_2 \cdot \gamma_{\text{con-t}} \odot \text{HypMLP}(Z_{\text{rc}}^t) + \\ \beta_3 \cdot \gamma_{\text{con-v}} \odot \text{HypMLP}(Z_{\text{rc}}^v), \end{aligned} \tag{12}$$

where β_1, β_2, and β_3 are learnable parameters that control the relative importance of each component.

2.5 Final Fusion and Classification

The final fusion combines features from different pathways denoted as F_U, F_C, and F_{rc} using a

weighted fusion mechanism, which is formulated as:

$$z_{\text{com}} = \text{WF}(F_U, F_C, F_{\text{rc}}) = \sum_{i=1}^{n} w_i \odot F_i, \quad (13)$$

Here, w_i are the learnable weights, and $\odot$ represents the element-wise multiplication. This integration ensures that the classification decision incorporates the full spectrum of semantic alignment patterns identified by our framework. Then we employ a hyperbolic-to-Euclidean transformation to project the final fusion features z_{com} to the Euclidean space, defined as:

$$z_{\text{euc}} = \text{HypToEuc}(z_{\text{com}}) = \text{MLP}(\log_{\mathbf{0}}(z_{\text{com}})), \quad (14)$$

where $\log_{\mathbf{0}}$ represents the logarithmic map from the hyperbolic space to the Euclidean space, and MLP is a multi-layer perceptron.

The final classification is performed by passing the projected features through a Multi-Layer Perceptron to the predicted label, defined as:

$$\hat{y} = \text{Softmax}(\text{MLP}(z_{\text{euc}})). \quad (15)$$

2.6 RumorCone Loss Function

The RumorCone framework is designed to model semantic consistency across hierarchical abstraction levels using hyperbolic geometry. To support its unique contributions, such as cross-modal alignment, geometry-aware learning, and hypothesis-driven reasoning, we formulate a composite loss function composed of four parts: classification loss, contrastive consistency loss, manifold-aware regularization, and a semantic consistency term.

Focal Supervised Classification Loss To address class imbalance and improve sensitivity to hard-to-classify samples, we apply a focal cross-entropy loss with class-balancing:

$$\mathcal{L}_{cls} = -\frac{1}{N} \sum_{i=1}^{N} \alpha_{y_i} (1 - p_{i,y_i})^{\gamma} \log(p_{i,y_i}), \quad (16)$$

Here, p_{i,y_i} is the predicted probability of the true class y_i, α_{y_i} is the class-specific weight, and γ is the focusing parameter that emphasizes harder examples.

Cross-Modal Contrastive Consistency Loss To preserve semantic consistency across modalities and their fused representations, we introduce a multi-part contrastive loss:

$$\mathcal{L}_{consist} = \mathcal{L}_{contr}(h^t, h^v) + \mathcal{L}_{contr}(h^t, z_{com}) + \mathcal{L}_{contr}(h^v, z_{com}), \quad (17)$$

where each term uses a temperature-scaled contrastive loss:

$$\mathcal{L}_{contr}(a, b) = -\frac{1}{N} \sum_{i=1}^{N} \log \frac{\exp(a_i^T b_i/\tau)}{\sum_{j=1}^{N} \exp(a_i^T b_j/\tau)}, \quad (18)$$

This loss aligns embeddings of corresponding textual, visual, and joint representations, ensuring consistent reasoning across abstraction levels.

Manifold-Aware Regularization To preserve the structure of hyperbolic space representations, we incorporate a geometry-aware loss that maintains Lorentzian manifold constraints and ensures meaningful embedding spread:

$$\mathcal{L}_{\text{manifold}} = \text{MSE}(\langle z_{\text{com}}, z_{\text{com}} \rangle_{\mathcal{L}}, -1) + \lambda_{\text{spread}} \cdot \frac{1}{N^2} \sum_{i \neq j} \exp(-\alpha \cdot d_{\mathbb{H}}(z_i, z_j)) \quad (19)$$

where $\langle \cdot, \cdot \rangle_{\mathcal{L}}$ is the Lorentzian inner product, $d_{\mathbb{H}}(z_i, z_j)$ is the hyperbolic distance between points z_i and z_j, α is a temperature-like scaling coefficient (default: $\alpha = 5$), λ_{spread} is a weighting term controlling the contribution of this regularizer, $\text{MSE}(\cdot, \cdot)$ ensures points lie on the Lorentz manifold.

Semantic Consistency Loss To reflect Hypothesis H1, stating that visual-semantic hierarchies become increasingly aligned or misaligned as abstraction levels deepen, we define a hyperbolic geometry-aware alignment loss. Let z_l^t, z_l^v denote the Lorentz-projected text and image embeddings at level $l \in \{\text{Base}, \text{Mid}, \text{High}\}$. Define:

$$\Delta_i = -d_{\mathbb{H}}(z_H^t, z_H^v) + d_{\mathbb{H}}(z_B^t, z_B^v), \quad (20)$$

The class-specific target slope t_i is:

$$t_i = \begin{cases} +1 & \text{if } y_i = 0 \quad \text{(non-rumor)} \\ -1 & \text{if } y_i = 1 \quad \text{(rumor)} \\ 0 & \text{if } y_i = 2 \quad \text{(unverified)} \end{cases} \quad (21)$$

The final semantic consistency loss is:

$$\mathcal{L}_{H1} = \frac{1}{N} \sum_{i=1}^{N} (\Delta_i - t_i)^2, \quad (22)$$

This loss encourages increasing alignment for non-rumors, increasing misalignment for rumors, and a neutral slope for unverified content, consistent with the semantic behaviors proposed in Hypothesis H1.

Total Loss. The total training objective for RumorCone is the weighted sum of the four components described above:

$$\begin{aligned}\mathcal{L}_{total} = \mathcal{L}_{cls} + \lambda_{consist} \cdot \mathcal{L}_{consist} \\ +\lambda_{manifold} \cdot \mathcal{L}_{manifold} + \lambda_{H1} \cdot \mathcal{L}_{H1},\end{aligned} \tag{23}$$

Here, $\lambda_{consist}$, $\lambda_{manifold}$, and λ_{H1} are scalar hyperparameters used to balance the contribution of each auxiliary objective relative to the supervised classification loss. These are tuned via validation to ensure stable convergence and generalization.

3 Experiment Settings and Results Analysis

3.1 Datasets and Baselines

Datasets. We evaluated the RumorCone framework on two multimodal rumor detection benchmark datasets: MR^2-E and MR^2-C, the multimodal multilingual retrieval-augmented dataset for rumor detection.

Baseline Methods. We compared RumorCone with state-of-the-art methods across text-only (BERT (Devlin et al., 2018), RoBERTa (Liu et al., 2019)), image-only (ResNet (He et al., 2015), Vision Transformer (Dosovitskiy et al., 2021)), and multimodal approaches (MVAE (Khattar et al., 2019), CLIP (Radford et al., 2021), HSEN (Zhang et al., 2023), HAGNN (Xu et al., 2023b), HMCAN (Qian et al., 2021), FSRU (Lao et al., 2024)), AAR (Zheng et al., 2025)). We comprehensively describe each baseline in Appendix D.

3.2 Evaluation Against Baseline Methods

To evaluate the effectiveness of the RumorCone framework, we conducted comprehensive experiments comparing it against several state-of-the-art rumor detection models. Table 1 presents the performance metrics across the MR^2-E and MR^2-C datasets, offering insights into how our hyperbolic geometry-based approach compares to existing methods.

RumorCone consistently outperforms all baseline methods, achieving an accuracy improvement of 4.61% (86.71% vs. 82.10%) and F1 score improvement of 4.77% (86.76% vs. 81.99%) over the strongest baseline (HMCAN) on the MR^2-E dataset. For the MR^2-C dataset, RumorCone demonstrates competitive performance with an F1 score of 85.48%, showing a 2.59% improvement over HMCAN (82.89%). Compared to the recently proposed FSRU and AAR models, which achieve F1 scores of 78.66% and 79.05% on MR^2-E respectively, RumorCone shows clear improvements of over 7% in F1 score.

The performance gap between RumorCone and flat-space models like CLIP (81.40% F1 on MR^2-E) validates our first hypothesis that posts maintaining high semantic consistency between image and caption across multiple abstraction levels are more accurately classified in hyperbolic space. This is further evidenced by our ablation study, where the Euclidean-only variant achieved only 81.55% F1 score compared to our full model. RumorCone's superior performance over hierarchical approaches like HSEN (80.37% F1) and HAGNN (81.37% F1) confirms our second hypothesis that explicit modeling of semantic inconsistencies across abstraction levels enhances rumor detection. The substantial performance gap between RumorCone and unimodal approaches (BERT: 77.57% F1, ResNet: 75.12% F1) underscores the importance of multimodal reasoning in rumor detection.

These results demonstrate that RumorCone's integration of hyperbolic geometry with hierarchical semantic alignment provides a principled approach to detecting the subtle inconsistencies characteristic of multimodal rumors, advancing the state-of-the-art in this critical application domain.

3.3 Ablation Studies

To systematically evaluate the contribution of each component in the RumorCone framework, we conducted an extensive ablation study. Table 2 presents the performance metrics across both datasets when removing or modifying key components of our model architecture.

Geometry Analysis. We first evaluated the impact of different geometric spaces on model performance. Removing the MERU hyperbolic-Euclidean fusion module resulted in a 3.05% decrease in F1 score on MR^2-E (86.76% vs. 83.71%), indicating the crucial role of our hybrid geometric approach. The Euclidean Only variant achieved an F1 score of 81.55%, and the Hyperbolic Only variant performed even worse at 70.38% F1, confirming that neither space alone

Table 1: Performance comparison of RumorCone with baseline models on MR^2-E and MR^2-C datasets. Some baselines are adapted from binary to multiclass for comparison.

Baselines	Method	MR^2-E				MR^2-C			
		Accuracy	Precision	Recall	F1	Accuracy	Precision	Recall	F1
Text-Based	BERT (Devlin et al., 2018)	78.34±1.23	78.69±1.11	77.02±1.01	77.57±1.13	82.42±1.07	79.21±0.98	77.45±0.86	78.05±1.18
	RoBERTa (Liu et al., 2019)	78.22±1.05	78.26±0.92	78.87±1.07	78.56±0.92	83.64±0.88	80.43±0.95	78.77±0.89	78.56±0.94
Image-Based	ResNet (He et al., 2015)	74.83±1.07	76.22±1.23	74.37±1.02	75.12±1.23	72.33±1.10	65.16±1.25	62.13±1.18	63.39±1.13
	Vision Transformer (Dosovitskiy et al., 2021)	72.14±1.29	74.33±1.28	72.29±1.36	73.28±1.33	70.75±1.26	64.54±1.24	62.24±1.14	64.89±1.23
Multimodal	MVAE (Khattar et al., 2019)	78.25±1.76	78.33±1.52	77.24±1.61	77.78±1.57	81.32±1.68	78.84±1.47	77.58±1.59	78.21±1.53
	CLIP (Radford et al., 2021)	81.29±1.22	82.05±1.27	81.06±1.32	81.40±1.22	83.87±1.25	81.77±1.31	81.29±1.23	80.54±1.29
	HSEN (Zhang et al., 2023)	80.14±1.11	81.29±1.08	79.45±1.02	80.37±1.07	84.23±1.01	82.05±1.15	80.71±1.09	81.33±1.05
	HAGNN (Xu et al., 2023b)	81.45±1.06	82.11±1.02	80.69±1.15	81.37±1.09	85.14±0.96	83.12±1.03	81.55±1.11	82.34±1.01
	HMCAN (Qian et al., 2021)	82.10±1.18	82.84±1.13	81.14±1.09	81.99±1.11	85.87±1.09	83.75±1.06	82.08±1.13	82.89±1.05
	FSRU (Lao et al., 2024)	80.31±0.40	77.85±0.35	79.50±0.38	78.66±0.37	79.42±0.36	76.92±0.32	78.67±0.35	77.78±0.33
	AAR (Zheng et al., 2025)	79.12±0.95	79.34±0.88	78.76±0.90	79.05±0.92	79.01±0.91	79.25±0.87	78.54±0.89	78.89±0.90
	RumorCone (Ours)	**86.71±1.17**	**86.82±1.24**	**86.71±1.19**	**86.76±1.22**	**85.45±1.14**	**85.62±1.18**	**85.34±1.15**	**85.48±1.17**

Table 2: Ablation study results for the RumorCone framework on MR^2-E and MR^2-C datasets.

Model	Variant	MR^2-E				MR^2-C			
		Accuracy	Precision	Recall	F1	Accuracy	Precision	Recall	F1
RumorCone	w/o MERU	83.36	83.17	84.25	83.71	82.18	81.95	83.04	82.49
	Hyperbolic Only (MERU)	82.42	75.29	66.07	70.38	81.17	73.46	64.92	68.95
	Euclidean Only (BERT+CLIP)	81.29	82.05	81.06	81.55	79.84	80.73	79.52	80.12
	No Cross-Modal	66.02	73.73	66.02	69.66	64.78	72.16	65.37	68.59
	No RumorCone pathways	63.98	67.47	63.98	65.68	62.24	65.89	62.73	64.25
	No Semantic Levels	78.25	78.33	77.24	77.78	76.93	77.12	75.98	76.54
	Full Model	**86.71**	**86.82**	**86.71**	**86.76**	**85.45**	**85.62**	**85.34**	**85.48**

can effectively capture the complex semantic hierarchies present in multimodal rumors. The substantial drop with hyperbolic-only embeddings suggests that while hyperbolic space efficiently represents hierarchical structures, it must be complemented by Euclidean space for optimal representation learning.

Semantic Hierarchy Analysis. The "No Semantic Levels" variant removed our hierarchical representation, resulting in a significant F1 score drop of 8.98% on MR^2-E (86.76% vs. 77.78%). This validates our hypothesis that explicitly modeling multiple semantic abstraction levels significantly enhances rumor detection performance. The hierarchical approach allows the model to identify inconsistencies that might be obscured at a single level of representation.

Cross-Modal Processing Analysis. The most dramatic performance decline was observed in the "No Cross-Modal" variant, with an F1 score of merely 69.66% on MR^2-E, representing a 17.10% decrease from the Full Model. This substantial gap demonstrates that cross-modal interactions are essential for identifying the semantic inconsistencies that characterize rumors. Similarly, the "No RumorCone pathways" variant performed even worse (65.68% F1), highlighting the critical importance of our proposed multi-pathway architecture for effective rumor detection.

These ablation results conclusively demonstrate that each component of the RumorCone framework contributes substantially to its overall performance. The hybrid geometric approach, multi-level semantic representations, and cross-modal pathways all play vital roles in capturing the subtle inconsistencies present in multimodal rumors. Particularly striking is the severe performance degradation when removing cross-modal processing components, which confirms that the ability to align and verify content across modalities is the cornerstone of effective rumor detection. The results empirically validate our theoretical hypothesis (H1) regarding the importance of hierarchical semantic alignment and the characteristic inconsistencies in rumors across abstraction levels.

3.4 Performance on Low-Resource and Multilingual Data

To assess generalization to low-resource and multilingual settings, we evaluated RumorCone on a newly compiled dataset of 100 real-world social media posts collected from reputable fact-checking platforms between 2024 - 2025. These posts span five major languages: English, Chinese,

Table 3: Performance of RumorCone and baselines on multilingual data with unified translation. Best results per row are bolded.

Language	Metric	MKV	FSRU	MVAE	HSEN	HAGNN	HMCAN	RumorCone
English	Accuracy	0.74	0.76	0.72	0.73	0.75	0.77	**0.79**
	F1 Score	0.70	0.72	0.68	0.69	0.71	0.72	**0.76**
Bengali	Accuracy	0.65	0.67	0.63	0.64	0.66	0.73	**0.76**
	F1 Score	0.60	0.63	0.58	0.60	0.62	**0.77**	0.73
Hindi	Accuracy	0.61	0.64	0.60	0.59	0.63	**0.77**	0.71
	F1 Score	0.56	0.59	0.55	0.54	0.58	0.61	**0.67**
Chinese	Accuracy	0.69	0.70	0.67	0.68	**0.72**	0.70	0.70
	F1 Score	**0.73**	0.71	0.69	0.69	0.70	0.72	0.68
Arabic	Accuracy	0.58	0.61	0.57	0.59	0.60	0.61	**0.67**
	F1 Score	0.55	0.60	0.54	0.56	0.58	0.65	**0.71**

Arabic, Bengali, and Hindi. Table 3 presents the detailed results for all baselines, including accuracy and F1 scores.

All models experienced performance degradation on this out-of-distribution dataset, particularly in low-resource languages such as Bengali and Hindi. Misclassifications were often caused by idiomatic or culturally nuanced expressions in the original captions, which became semantically ambiguous after translation. This semantic abstraction drift at the mid-level led to inconsistencies in cross-modal alignment and degraded prediction performance. Applying unified translation before hierarchical processing substantially improves performance in low-resource languages, demonstrating the robustness of RumorCones hierarchical, cross-modal semantic reasoning even in challenging multilingual settings.

These results demonstrate that RumorCone maintains competitive performance across multilingual and low-resource settings. Unified translation before hierarchical semantic abstraction mitigates semantic drift and enhances cross-modal alignment, particularly for languages where idiomatic or culturally specific expressions are prevalent.

3.5 Robustness to Partial or Noisy Modalities

To investigate RumorCone's behavior under incomplete or noisy multimodal inputs, we conducted experiments on the MR^2-E dataset with randomly masked modalities. Specifically, 30% of images or text inputs were masked to simulate real-world degradation. Table 4 presents the resulting performance metrics.

While performance drops with partial input, the model remains reasonably effective. This is consistent with our ablation study showing the No Cross-Modal variant leads to a large F1 drop (69.66%, Table 2). RumorCones modality-specific pathways and consistency-aware modulation dynamically weight each modality based on reliability, mitigating but not fully eliminating the impact of missing data.

Table 4: RumorCone performance under partial modality masking on MR^2-E dataset.

Condition	Acc.(%)	Prec.(%)	Rec.(%)	F1(%)
Full Model	86.71	86.82	86.71	86.76
30% Text Masked	80.03	79.94	79.32	79.96
30% Image Masked	81.28	80.61	80.85	81.18

4 Conclusion

In this paper, we presented the RumorCone framework, a novel approach to multimodal rumor detection that leverages hierarchical semantic alignment and hyperbolic geometry. The primary objective of RumorCone is to address the challenges posed by the complex nature of multimodal content, particularly the misalignment between textual and visual features that characterize many rumors. By modeling these semantic relationships across multiple abstraction levels, the framework improves the ability to detect subtle inconsistencies inherent in deceptive content. We introduced and thoroughly evaluated RumorCone on two benchmark datasets, MR^2-E and MR^2-C, which incorporate both textual and visual content from various social media platforms. Our experiments demonstrated that RumorCone outperforms state-of-the-art methods in multimodal rumor detection, achieving superior performance across multiple evaluation metrics.

Limitations

The RumorCone framework, though effective in modeling hierarchical semantic graphs, is also severely lacking. Its performance is heavily dependent on having good quality translation in multilingual settings. In experiments, it reported a sharp drop in accuracy for low-resource languages without translation, indicative of its reliance on semantic normalization. Although hyperbolic geometry is useful for preserving abstraction and hierarchy, it has the drawback of optimization complexity and scalability, particularly on large, noisy, or imbalanced datasets.

Another key limitation comes from the incomplete modality problem. In real-world social media posts, one or more of the modalities (e.g., image, or text) would be missing or corrupted. RumorCone's performance degrades against such partial inputs, and that implies the need for more robust modality-invariant reasoning techniques. Further, low-resource settings, especially under-represented languages, are both data-poor and domain-mismatch issues since current multilingual rumor corpora often skew heavily in favor of English, and Chinese, leaving most world languages poorly supported.

Finally, RumorCone lacks external fact-checking or real-world factual verification system integration, limiting its ability to anchor rumors in verified sources. Although it captures rumors through semantic and contextual signals, it fails to cross-check against verified knowledge bases or real-time fact stores, which precludes its use in high-stakes scenarios where factual accuracy is crucial.

Acknowledgement

This work is supported by the National Natural Science Foundation of China (No. 62372043, 62102026).

References

Rasha M. Albalawi, Amani T. Jamal, Alaa O. Khadidos, and Areej M. Alhothali. 2023. Multimodal Arabic Rumors Detection. *IEEE Access*, 11:9716–9730.

Abderrazek Azri, Cécile Favre, Nouria Harbi, Jérôme Darmont, and Camille Noûs. 2021. Calling to CNN-LSTM for Rumor Detection: A Deep Multi-channel Model for Message Veracity Classification in Microblogs. ArXiv: 2110.15727 [cs].

Christina Boididou, Symeon Papadopoulos, Yiannis Kompatsiaris, Steve Schifferes, and Nic Newman. 2014. Challenges of computational verification in social multimedia. In *Proceedings of the 23rd International Conference on World Wide Web*, WWW '14 Companion, page 743748, New York, NY, USA. Association for Computing Machinery.

Jiaxin Chen, Zekai Wu, Zhenguo Yang, Haoran Xie, Fu Lee Wang, and Wenyin Liu. 2021. Multimodal Fusion Network with Latent Topic Memory for Rumor Detection. In *2021 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE. Place: Shenzhen, China.

Karan Desai, Maximilian Nickel, Tanmay Rajpurohit, Justin Johnson, and Ramakrishna Vedantam. 2023. Hyperbolic Image-Text Representations. In *Proceedings of the International Conference on Machine Learning*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Bhuwan Dhingra, Christopher J. Shallue, Mohammad Norouzi, Andrew M. Dai, and George E. Dahl. 2018. Embedding text in hyperbolic spaces. *Preprint*, arXiv:1806.04313.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. *Preprint*, arXiv:2010.11929.

Boyang Fu and Jie Sui. 2022. Multi-modal affine fusion network for social media rumor detection. *PeerJ Computer Science*, 8:e928.

Yadong Gu, Mijit Ablimit, and Askar Hamdulla. 2023. Fake News Detection Based on Cross-modal Co-attention. In *2023 IEEE 4th International Conference on Pattern Recognition and Machine Learning (PRML)*, pages 401–406. IEEE. Place: Urumqi, China.

Hao Guo, Jiuyang Tang, Weixin Zeng, Xiang Zhao, and Li Liu. 2021. Multi-modal entity alignment in hyperbolic space. *Preprint*, arXiv:2106.03619.

Yunhui Guo, Xudong Wang, Yubei Chen, and Stella X. Yu. 2022. Clipped hyperbolic classifiers are super-hyperbolic classifiers. *Preprint*, arXiv:2107.11472.

Zhiwei Guo, Zhenguo Yang, and Dahuang Liu. 2024. Rumor detection based on cross-modal information-enhanced fusion network. In *2024 16th International Conference on Advanced Computational Intelligence (ICACI)*, pages 158–164.

H. Han, Z. Ke, X. Nie, L. Dai, and W. Slamu. 2023a. Multimodal fusion with dual-attention based on textual double-embedding networks for rumor detection. *Applied Sciences*, 13(8):4886.

Huawei Han, Jianlei Yang, and Wushour Slamu. 2023b. Cascading Modular Multimodal Cross-attention Network for Rumor Detection. In *2023 IEEE International Conference on Control, Electronics and Computer Technology (ICCECT)*, pages 974–980. IEEE. Place: Jilin, China.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Deep residual learning for image recognition. *Preprint*, arXiv:1512.03385.

Xuming Hu, Zhijiang Guo, Junzhe Chen, Lijie Wen, and Philip S. Yu. 2023. MR2: A Benchmark for Multimodal Retrieval-Augmented Rumor Detection in Social Media. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2901–2912. ACM. Place: Taipei Taiwan.

Fangting Jiang, Gang Liang, Jin Yang, and Liangyin Chen. 2023. MMRDF: An improved multitask multimodal rumor detection framework. *Electronics Letters*, 59(10):e12811.

Zhiwei Jin, Juan Cao, Han Guo, Yongdong Zhang, and Jiebo Luo. 2017. Multimodal fusion with recurrent neural networks for rumor detection on microblogs. pages 795–816.

Dhruv Khattar, Jaipal Singh Goud, Manish Gupta, and Vasudeva Varma. 2019. Mvae: Multimodal variational autoencoder for fake news detection. In *The World Wide Web Conference*, WWW '19, page 29152921, New York, NY, USA. Association for Computing Machinery.

Valentin Khrulkov, Leyla Mirvakhabova, Evgeniya Ustinova, Ivan Oseledets, and Victor Lempitsky. 2020. Hyperbolic image embeddings. *Preprint*, arXiv:1904.02239.

Wonjae Kim, Sanghyuk Chun, Taekyung Kim, Dongyoon Han, and Sangdoo Yun. 2024. Hype: Hyperbolic entailment filtering for underspecified images and texts. *Preprint*, arXiv:2404.17507.

Fanjie Kong, Yanbei Chen, Jiarui Cai, and Davide Modolo. 2024. Hyperbolic learning with synthetic captions for open-world detection. *Preprint*, arXiv:2404.05016.

Hyeongjun Kwon, Jinhyun Jang, Jin Kim, Kwonyoung Kim, and Kwanghoon Sohn. 2024. Improving visual recognition with hyperbolical visual hierarchy mapping. *Preprint*, arXiv:2404.00974.

An Lao, Qi Zhang, Chongyang Shi, Longbing Cao, Kun Yi, Liang Hu, and Duoqian Miao. 2024. Frequency spectrum is more effective for multimodal representation and fusion: a multimodal spectrum rumor detector. In *Proceedings of the Thirty-Eighth AAAI Conference on Artificial Intelligence and Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence and Fourteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'24/IAAI'24/EAAI'24. AAAI Press.

X. Li, J. Wang, and H. Zhang. 2024a. Cross-modality integration framework with prediction, perception, and discrimination for video anomaly detection. *IEEE Transactions on Multimedia*, 26(4):2235–2247.

Yang Li, Liguang Liu, Jiacai Guo, Lap-Kei Lee, Fu Lee Wang, and Zhenguo Yang. 2024b. Mkv: Mapping key semantics into vectors for rumor detection. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '24, page 25122516, New York, NY, USA. Association for Computing Machinery.

Shaoteng Liu, Jingjing Chen, Liangming Pan, Chong-Wah Ngo, Tat-Seng Chua, and Yu-Gang Jiang. 2020. Hyperbolic visual embedding learning for zero-shot recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

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. *Preprint*, arXiv:1907.11692.

Jiandong Lv, Xingang Wang, and Cuiling Shao. 2023. TMIF: transformer-based multi-modal interactive fusion for automatic rumor detection. *Multimedia Systems*, 29(5):2979–2989.

Zhong Nanjiang, Zhou Guomin, Ding Weijie, and Zhang Jiawen. 2022. A Rumor Detection Method Based on Multimodal Information Fusion. In *2022 IEEE 5th International Conference on Electronics Technology (ICET)*, pages 1032–1037. IEEE. Place: Chengdu, China.

Maximilian Nickel and Douwe Kiela. 2017. Poincaré embeddings for learning hierarchical representations. *Preprint*, arXiv:1705.08039.

Yucai Pang, Xuehong Li, Shihong Wei, Qian Li, and Yunpeng Xiao. 2024. Topic to Image: A Rumor Detection Method Inspired by Image Forgery Recognition Technology. *IEEE TRANSACTIONS ON COMPUTATIONAL SOCIAL SYSTEMS*, 11(2).

Shengsheng Qian, Jinguang Wang, Jun Hu, Quan Fang, and Changsheng Xu. 2021. Hierarchical multi-modal contextual attention network for fake news detection. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '21, page 153162, New York, NY, USA. Association for Computing Machinery.

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 *International Conference on Machine Learning*, pages 8748–8763. PMLR.

Hongyan Ran and Caiyan Jia. 2023. Unsupervised Cross-Domain Rumor Detection with Contrastive Learning and Cross-Attention. *arXiv preprint*. ArXiv:2303.11945 [cs].

Ekraam Sabir, Wael AbdAlmageed, Yue Wu, and Prem Natarajan. 2018. Deep Multimodal Image-Repurposing Detection. In *Proceedings of the 26th ACM international conference on Multimedia*, pages 1337–1345. ACM. Place: Seoul Republic of Korea.

Sudhakar Sengan, Subramaniyaswamy Vairavasundaram, Logesh Ravi, Ahmad Qasim Mohammad Al-Hamad, Hamzah Ali Alkhazaleh, and Meshal Alharbi. 2024. Fake News Detection Using Stance Extracted Multimodal Fusion-Based Hybrid Neural Network. *IEEE TRANSACTIONS ON COMPUTATIONAL SOCIAL SYSTEMS*, 11(4).

Aditya Sinha, Siqi Zeng, Makoto Yamada, and Han Zhao. 2024. Learning structured representations with hyperbolic embeddings. *Preprint*, arXiv:2412.01023.

Chenguang Song, Nianwen Ning, Yunlei Zhang, and Bin Wu. 2021. A multimodal fake news detection model based on crossmodal attention residual and multichannel convolutional neural networks. *Information Processing & Management*, 58(1):102437.

Mengzhu Sun, Xi Zhang, Jianqiang Ma, Sihong Xie, Yazheng Liu, and Philip S. Yu. 2023a. Inconsistent matters: A knowledge-guided dual-consistency network for multi-modal rumor detection. *IEEE Trans. on Knowl. and Data Eng.*, 35(12):1273612749.

Tiening Sun, Zhong Qian, Peifeng Li, and Qiaoming Zhu. 2023b. Graph Interactive Network with Adaptive Gradient for Multi-Modal Rumor Detection. In *Proceedings of the 2023 ACM International Conference on Multimedia Retrieval*, pages 316–324. ACM. Place: Thessaloniki Greece.

Y. Tian, F. Gao, and C. Zhang. 2020. Deep supervised multimodal semantic autoencoder for cross-modal retrieval. *Pattern Recognition*, 108:107569.

Ge Wang, Li Tan, Ziliang Shang, and He Liu. 2023a. Multimodal dual emotion with fusion of visual sentiment for rumor detection. *Multimed Tools Appl*.

Hejian Wang, Peng Chen, Qinglei Guo, Lihua Zhao, Xiaoxiao Ma, Xuefeng Li, Bo Gao, and Yongliang Li. 2023b. False Information Detection Based on Multimodal Attention Detection Network. In *2023 6th International Conference on Artificial Intelligence and Big Data (ICAIBD)*, pages 530–534. IEEE. Place: Chengdu, China.

Jenq-Haur Wang, Mehdi Norouzi, and Shu Ming Tsai. 2022. Multimodal Content Veracity Assessment with Bidirectional Transformers and Self-Attention-based Bi-GRU Networks. In *2022 IEEE Eighth International Conference on Multimedia Big Data (BigMM)*, pages 133–137. IEEE. Place: Naples, Italy.

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*, KDD '18, page 849857, New York, NY, USA. Association for Computing Machinery.

Zhuang Wang and Jie Sui. 2021. Multilevel Attention Residual Neural Network for Multimodal Online Social Network Rumor Detection. *Front. Phys.*, 9:711221.

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, Online. Association for Computational Linguistics.

Fan Xu, Pinyun Fu, Qi Huang, Bowei Zou, AiTi Aw, and Mingwen Wang. 2023a. Leveraging contrastive learning and knowledge distillation for incomplete modality rumor detection. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 13492–13503, Singapore. Association for Computational Linguistics.

Shouzhi Xu, Xiaodi Liu, Kai Ma, Fangmin Dong, Basheer Riskhan, Shunzhi Xiang, and Changsong Bing. 2023b. Rumor detection on social media using hierarchically aggregated feature via graph neural networks. *Appl Intell*, 53(3):3136–3149.

Facheng Yan, Mingshu Zhang, Bin Wei, Kelan Ren, and Wen Jiang. 2024. Sard: Fake news detection based on clip contrastive learning and multimodal semantic alignment. *Journal of King Saud University - Computer and Information Sciences*, 36(8):102160.

Yang Yang, Ran Bao, Weili Guo, De-Chuan Zhan, Yilong Yin, and Jian Yang. 2023. Deep visual-linguistic fusion network considering cross-modal inconsistency for rumor detection. *Sci. China Inf. Sci.*, 66(12):222102.

Long Ying, Hui Yu, Wang Jinguang, Yongze Ji, and Shengsheng Qian. 2021. Multi-level multi-modal cross-attention network for fake news detection. *IEEE Access*, PP:1–1.

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*, MM '19, page 19421951, New York, NY, USA. Association for Computing Machinery.

Huaiwen Zhang, Shengsheng Qian, Quan Fang, and Changsheng Xu. 2021. Multimodal Disentangled Domain Adaption for Social Media Event Rumor Detection. *IEEE Trans. Multimedia*, 23:4441–4454.

Qiang Zhang, Jiawei Liu, Fanrui Zhang, Jingyi Xie, and Zhengjun Zha. 2023. Hierarchical semantic enhancement network for multimodal fake news detection. *ACM Multimedia*.

Y. Zhang, L. Chen, and P. Wu. 2024. Multi-modal semantic understanding with contrastive cross-modal feature alignment. *Artificial Intelligence Review*, 57(3):1567–1584.

Xiaofan Zheng, Minnan Luo, and Xinghao Wang. 2025. Unveiling fake news with adversarial arguments generated by multimodal large language models. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 7862–7869, Abu Dhabi, UAE. Association for Computational Linguistics.

Nanjiang Zhong, Guomin Zhou, Weijie Ding, and Jiawen Zhang. 2022. A Rumor Detection Method Based on Multimodal Feature Fusion by a Joining Aggregation Structure. *Electronics*.

Honghao Zhou, Tinghuai Ma, Huan Rong, Yurong Qian, Yuan Tian, and Najla Al-Nabhan. 2022. MDMN: Multi-task and Domain Adaptation based Multi-modal Network for early rumor detection. *Expert Systems with Applications*, 195:116517.

Wei Zhou, Chenghao Li, Lin Zuo, Min Gao, and Junhao Wen. 2025. Double chain multimodal feature learning algorithm for rumor detection via contrastive learning. *SSRN*.

Xinyi Zhou, Jindi Wu, and Reza Zafarani. 2020a. Multimodal learning for social media rumor detection. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, pages 1445–1454.

Xinyi Zhou, Jindi Wu, and Reza Zafarani. 2020b. Safe: Similarity-aware multi-modal fake news detection. In *WWW*.

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.

Ting Zou, Zhong Qian, Peifeng Li, and Qiaoming Zhu. 2023. Cross-Modal Adversarial Contrastive Learning for Multi-Modal Rumor Detection. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5, Rhodes Island, Greece. IEEE.

A Visualization of hierarchical semantic processing

B Key Functions in RumorCone Method

The Lorentzian exponential mapping $\exp_{\mathbf{0}}$ is defined as:

$$\exp_{\mathbf{0}}(x) = \left(\sqrt{\frac{1}{c} + \|\lambda x\|^2},\ \frac{\sinh(\sqrt{c}\|x\|)}{\sqrt{c}\|x\|} x \right). \tag{24}$$

where spatial embeddings are scaled using hyperbolic sine to ensure correct curvature-based projection.

The hyperbolic attention mechanism is defined as:

$$\text{Attention}(Q, K, V) = \bigoplus_{j=1}^{n} \alpha_j \otimes_{\mathbb{H}} v_j \tag{25}$$

$$\alpha_j = \frac{\exp(-\tau \cdot d_{\mathbb{H}}(q, k_j))}{\sum_{l=1}^{n} \exp(-\tau \cdot d_{\mathbb{H}}(q, k_l))} \tag{26}$$

where $\bigoplus$ represents iterative Möbius addition, $\otimes_{\mathbb{H}}$ is Möbius scalar multiplication, and τ is a temperature parameter controlling the softness of attention weights.

The hyperbolic distance $d_{\mathbb{H}}(q, k)$ between two vectors q and k is calculated by using the Lorentzian metric, defined as:

$$d_{\mathbb{H}}(q, k) = \text{arcosh}\left(1 + 2\frac{\|q - k\|^2}{(1 - \|q\|^2)(1 - \|k\|^2)}\right) \tag{27}$$

C Datasets

The MR^2 (Hu et al., 2023) dataset includes rumors with both images and texts, providing evidence from both modalities retrieved from the internet. It is divided into three categories: Content-based Method, Propagation-based Method, and Retrieval-based Method. In this study, we focused on the Content-based Method. To address class imbalance, we employed weighted sampling during training to enhance model generalization across all classes. The dataset MR^2-E and MR^2-C, sourced from Twitter and Weibo, respectively.

D Baselines

We compared RumorCone against several state-of-the-art methods across three categories:

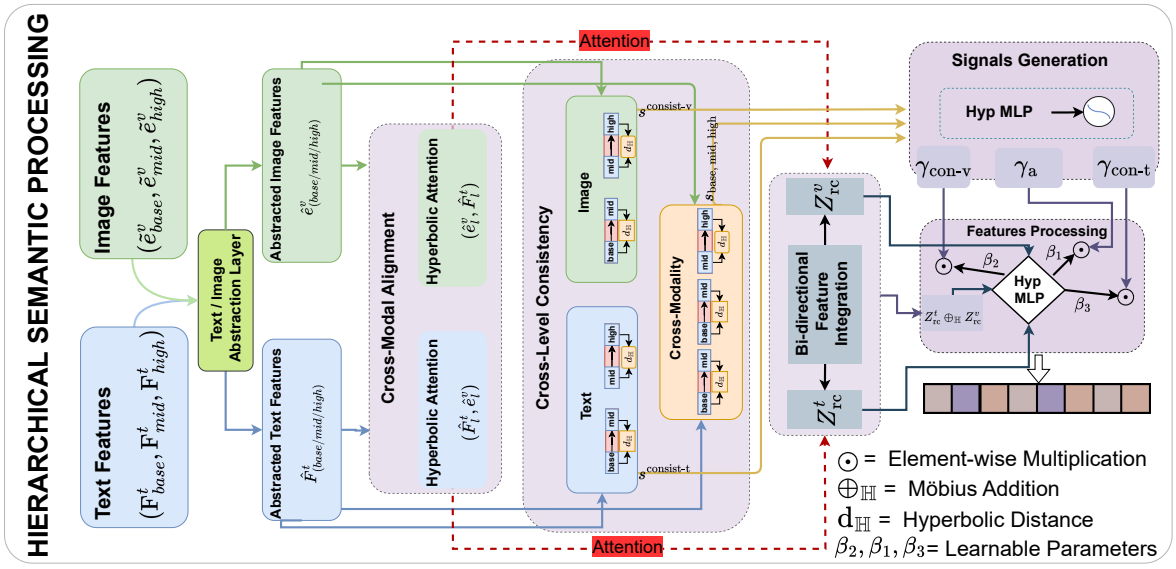

Figure 3: Visualization of hierarchical semantic processing mechanisms in RumorCone. The diagram illustrates: (A) Within-level alignment between text and image at each abstraction level, (B) Cross-level consistency checks within each modality and (C) RumorCone feature generation

1. **Text-Only Methods:** Text-based approaches rely solely on linguistic features for rumor detection. BERT (Devlin et al., 2018) leverages bidirectional transformer encoders pre-trained on massive text corpora to extract contextual representations, enabling it to capture semantic nuances and deceptive patterns in textual content. RoBERTa (Liu et al., 2019), an optimized extension of BERT with improved pre-training methodology, offers enhanced performance through dynamic masking and larger batch sizes. While these models achieve reasonable performance by identifying linguistic deception signals, they fundamentally cannot detect inconsistencies between text and corresponding images, a common characteristic of multimodal rumors. Our experiments confirm this limitation, as text-only approaches miss crucial visual evidence that often contradicts false textual claims.

2. **Image-Only Methods:** Visual-based approaches focus exclusively on image manipulation detection. ResNet (He et al., 2015) employs deep residual learning to extract hierarchical visual features, enabling the identification of subtle visual manipulation artifacts common in misleading content. Vision Transformer (Dosovitskiy et al., 2021) adapts the transformer architecture to computer vision by processing images as sequences of patches, effectively modeling long-range dependencies in visual data that may indicate tampering. However, these approaches struggle to contextualize images within their accompanying textual claims, leading to high false positive rates when legitimate images are presented with misleading captions. Without cross-modal reasoning capabilities, image-only methods demonstrate significantly lower performance than multimodal approaches in our experiments.

3. **Multimodal Methods:** Recent approaches leverage both textual and visual information for more comprehensive rumor detection. MVAE (Khattar et al., 2019) employs a multimodal variational autoencoder to learn joint representations of text and images, significantly improving detection reliability over unimodal approaches. CLIP (Radford et al., 2021) uses contrastive learning to align visual and language representations from 400 million image-text pairs, enabling zero-shot transfer to rumor detection tasks. More specialized architectures include HSEN (Zhang et al., 2023), which enhances semantic alignment across modalities through hierarchical feature extraction, and HAGNN (Xu et al., 2023b), which models rumor propagation using graph neural networks

with attention mechanisms. HMCAN (Qian et al., 2021) combines multi-modal context information and hierarchical textual semantics, using BERT for text and ResNet for images fused through a multi-modal attention network. FSRU (Lao et al., 2024) further advances multimodal fusion by transforming textual and visual features into the frequency domain and applying cross-modal spectrum co-selection for more discriminative representations. Most recently, AAR (Zheng et al., 2025) introduces adversarial arguments generated by multimodal large language models to guide cross-attentional reasoning, achieving state-of-the-art results in fake news detection. While these approaches recognize the importance of cross-modal reasoning, they operate in Euclidean space, limiting their ability to capture the hierarchical semantic relationships that RumorCone explicitly models in hyperbolic space.

E Implementation details

E.1 Base, Mid and High-level abstraction for Text

To support RumorCones hierarchical reasoning mechanism, we generated two additional abstraction levels, Mid-level and High-level textual inputs, for each sample in the dataset, alongside the original (Base-level) text. The abstraction process is performed using a sentence-ranking approach based on contextual centrality, computed using embeddings from a multilingual BERT model. Specifically, for each input text, we first perform language-aware sentence segmentation. Then, sentence embeddings are extracted using the average of the last hidden state from BERT. A document embedding is computed as the mean of all sentence embeddings, and sentence importance scores are derived using cosine similarity with this document embedding. Sentences are then ranked by centrality and iteratively selected until a target length threshold is met 50% of the original for Mid-level abstraction, and 25% for High-level abstraction.

E.2 Hyperparameters for loss function

To ensure balanced optimization of RumorCones multi-objective architecture, we empirically tuned the loss weight hyperparameters using validation performance on the MR^2-E development set. Unless otherwise stated, the final values used across all experiments were: $\lambda_{consist} = 1.0$, $\lambda_{manifold} = 0.1$, and $\lambda_{H1} = 1.0$. The classification loss $\mathcal{L}cls$ was used as the primary supervision signal and therefore kept unweighted. We found that these values provided a stable trade-off between model performance and convergence speed, while avoiding overfitting to any single objective. The internal weight λ_{spread} in $\mathcal{L}_{manifold}$ was fixed to 0.5, following prior work on hyperbolic regularization.

E.3 Model Architecture

The RumorCone architecture integrates pre-trained modality-specific encoders (BERT-base-multilingual-cased for text of size 768 and ViT-Base-Patch16-224 for images), a MERU-based hyperbolic fusion module with Transformer L12_W512 text encoder and 512-size embedding, hierarchical semantic layers of 3 levels (base, mid, high) and 512-size features, and a cross-modal integration module utilizing multi-head self-attention (8 heads, 512-dim features). The fusion module projects features to a 64-dim hidden space and produces 3-class outputs. AdamW with component-specific learning rates (5e-6 for MERU, 2e-5 for main modules, 1e-5 for fusion), weight decay of 0.2, and a one-cycle schedule with cosine annealing, 15% warmup, and 1000 $\times$ decay are utilized for training. Mixed-precision training is performed with batch size 32 for a maximum of 100 epochs using early stopping (patience=10). The total loss includes focal cross-entropy (weight=2.0), adaptive class weights, contrastive consistency loss (weight=0.5, temperature=0.07), cross-modal contrastive loss (weight=0.2), hierarchical consistency (weight=0.3), and Lorentz manifold-aware regularization (weight=0.3), each balanced with learnable weights.

E.4 Evaluation Metrics

We report accuracy, precision, recall, and macro F1 score as our primary evaluation metrics, with macro F1 being especially important due to class imbalance. All experiments were conducted on NVIDIA GPUs with a PyTorch implementation.

F Semantic Consistency and Misalignment Patterns in RumorCone

In this section, we analyze the relationship between semantic consistency and misalignment patterns across different levels of abstraction in mul-

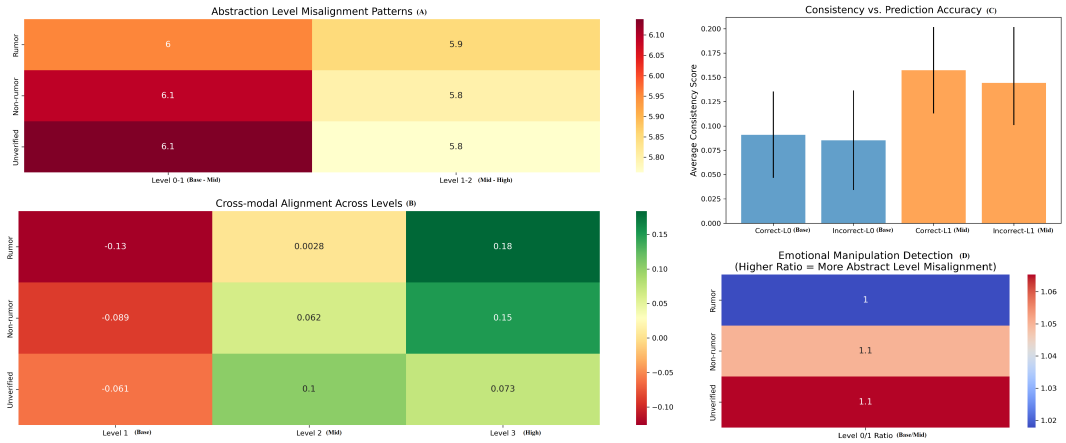

Figure 4: Cross-Modal Alignment scores for RumorCone and baseline models. The plot shows how well the textual and visual features align across different abstraction levels.

timodal rumor detection. The visualizations provide key insights into how misalignment between text and image impacts prediction accuracy, particularly in distinguishing between Rumor, Non-rumor, and unverified content. These findings align with the hypotheses proposed in the introduction 1 and support the need for models like RumorCone that can effectively handle complex semantic alignments across various abstraction levels.

The Abstraction Level Misalignment Patterns heatmap in Fig. 4 (A) shows that unverified content exhibits the highest misalignment at the base-mid level (6.1), followed by rumor and non-rumor content, while rumors, non-rumors, and unverified content exhibit low misalignment at the mid-high level. This suggests that these visual semantic contents are more misaligned at lower abstraction levels, while increased alignment at high abstraction levels indicates reduced ambiguity in the content from lower to higher levels. Our hypothesis (H1) states that posts with greater semantic alignment across abstraction levels help classify rumors. This observation aligns with the idea that rumors involve emotional manipulation, leading to greater misalignment between text and image.

The Cross-modal alignment across three hierarchical levels heatmap in Fig. 4 (B) further supports Hypothesis (H1), showing that Rumor content exhibits negative alignment at Base (-0.13) but improves at Mid (0.0028) and High (0.18) levels. This suggests that rumors, often emotionally manipulative, show greater misalignment at Base, where the content is more emotionally charged, and better alignment at higher abstraction levels, where more context and detail are captured. These findings suggest that semantic consistency increases with higher representational levels and correlates positively with prediction accuracy, potentially indicating that more coherent internal representations emerge in deeper processing layers. Particularly noteworthy in the emotional manipulation detection heatmap in Fig. 4 (D) shows a higher misalignment ratio for Rumor content (1.06 at Base/Mid ratio), further supporting the idea that rumors involve greater semantic inconsistency. These findings highlight the importance of addressing misalignments across abstraction levels to improve further rumor detection and support the development of frameworks.

The consistency vs. prediction accuracy plot in Fig. 4 (C) further emphasizes the importance of semantic alignment in improving prediction accuracy. Correct predictions consistently show higher average consistency scores compared to incorrect ones, highlighting the role of better alignment in enhancing model performance. These insights underscore the need for advanced frameworks like RumorCone, which leverages hierarchical semantic alignment and hyperbolic geometry to capture the complex relationships between text and image. By addressing these semantic misalignments, RumorCone improves its ability to detect rumors, particularly those involving emotional manipulation, and provides a more effective approach for multimodal rumor detection.

G Confusion Matrix Analysis

To evaluate the classification performance, we present a confusion matrix in Fig. 5 that shows how well RumorCone distinguishes between ru-

mors, non-rumors, and unverified content. The confusion matrix highlights the true positives, false positives, true negatives, and false negatives, providing a clear picture of the models effectiveness in classifying the three categories. The confusion matrices for MR^2-C and MR^2-E both show strong performance in classifying Rumor, Non-rumor, and Unverified posts, with MR^2-E achieving an overall accuracy of 86.71% and MR^2-C achieving 85.45%. MR^2-E demonstrates excellent classification of Rumor posts with 87% accuracy, while MR^2-C excels even further, classifying 95% of Rumor posts correctly. Both models also show strong performance for Unverified content, with MR^2-E achieving 92% and MR^2-C achieving 82%. This indicates that both models are highly effective at identifying rumors and unverified content.

However, both models also show some misclassifications, particularly between Non-rumor and Unverified categories. MR^2-E misclassified 16% of Non-rumor posts as Unverified, while MR^2-C misclassified 14% of Unverified posts as Non-rumor. Additionally, MR^2-E has a higher rate of misclassifying Non-rumor posts as Rumors (5%) compared to MR^2-C (8%), which also misclassifies Non-rumors as Rumors but at a lower rate. Despite these errors, the overall classification accuracy remains high for both models.

In conclusion, bothMR^2-C andMR^2-E demonstrate robust performance in detecting Rumor and Non-rumor posts, with MR^2-C showing slightly better performance in the Rumor category. The misclassifications between Non-rumor and Unverified content highlight areas for improvement, but overall, both models perform well across all categories. With further refinements in distinguishing Non-rumor from Unverified content, both models could achieve even higher classification accuracy.

H 3D Visualization of Fused Embeddings in Hyperbolic Space

The 3D scatter plots in Fig. 6 for MR^2-E and MR^2-C both illustrate the fused embeddings of the final model's output for three classes: Class 0 (Blue), Class 1 (Red), and Class 2 (Green). In the plot for MR^2-E, the classes are clearly separated, with minimal overlap between the clusters, indicating that the model effectively distinguishes between the classes in the fused embedding space. The distinct separation of the clusters suggests that MR^2-E has successfully captured the underlying structure of the data across the modalities, resulting in high classification performance. As seen in our interactive 3D visualization, this minimal overlap makes it clear that the classes are well differentiated, something that would be more challenging to observe in a 2D space. In contrast, the plot for MR^2-C shows considerable overlap between the classes, particularly in the regions where Class 1 (Red) and Class 2 (Green) intersect. This suggests that MR^2-C may not have as distinct class separation as MR^2-E, indicating that the model's feature fusion or embedding learning could benefit from further refinement. The overlap implies that MR^2-C may struggle more with distinguishing between certain classes, potentially requiring improved feature representation or clustering techniques. Overall, while MR^2-E shows strong class differentiation, MR^2-C may need further optimization to enhance its ability to separate the classes effectively. We have identified in interactive 3D visualization, effectively highlights the differences for MR^2-C, demonstrating that the minimal overlap in the 3D space provides clearer class separation, a distinction that would be more challenging to identify in a 2D representation.

I Case Study

We examine four representative cases from two different settings (MR^2-E for English content, MR^2-C for Chinese content) to understand how the multimodal model processes and classifies social media posts that pair images with text. In Fig. 7 (Left), the post shows a cheerful image of Mario, paired with a caption admiring the character design. The model accurately classifies this as Unverified, assigning a high score of 0.58. Despite the harmless tone, the fictional nature of the image and non-factual caption likely contributed to the models uncertainty. This highlights the models sensitivity to semantic context and verifiability. Fig. 7 (Right) presents a post featuring Garth Brooks and a vague statement. Although the ground truth is Non-rumor, the model incorrectly predicts Unverified with a slightly higher confidence (0.45). The ambiguity of the caption and lack of concrete factual content may have confused the model, indicating potential limitations in understanding metaphor or sarcasm. As shown in Fig. 8 (Left), the Chinese post combines a meme image with an exaggerated caption about being

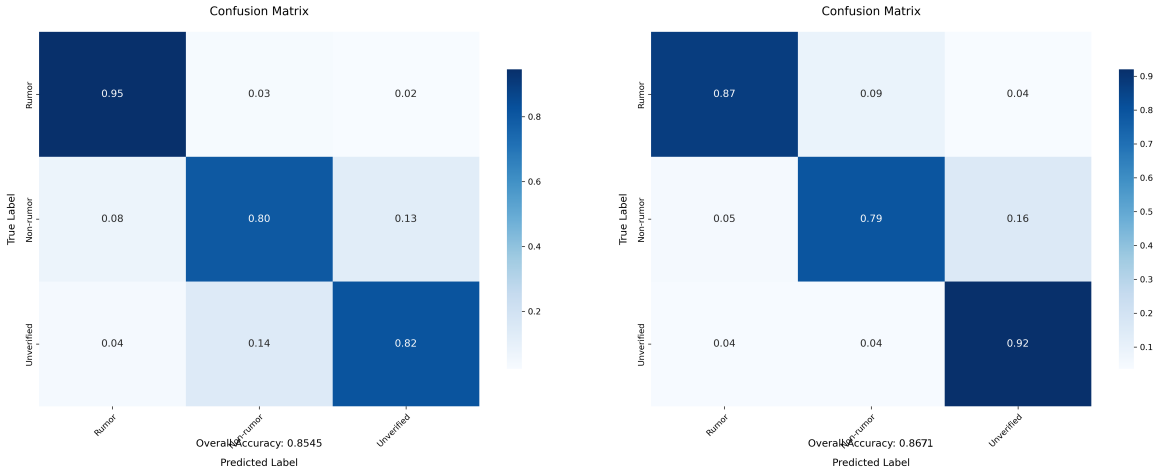

Figure 5: Confusion matrix for MR^2-C (Left) andMR^2-E (Right).

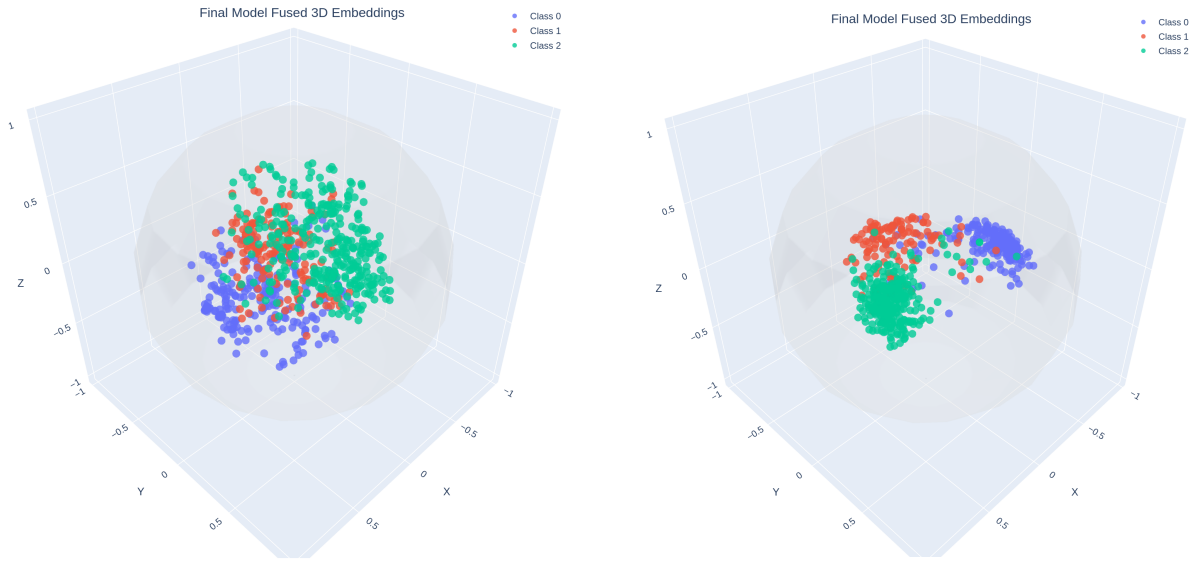

Figure 6: Best angle from the interactive 3D visualization of the fused embeddings in hyperbolic space for MR^2-C (Left) and MR^2-E (Right).

treated to drinks and meals. The model correctly labels this as Unverified with a high score of 0.65, recognizing the informal tone and unverifiable nature of the claim. This case shows the models ability to detect humor and exaggeration even in a non-English context. Fig. 8 (Right) depicts a serious post reporting radioactive contamination, paired with a formal image of a government official. The model classifies it as a Rumor with high confidence (0.75), aligning with the ground truth. The linguistic framing and alarming subject matter likely contributed to this classification. This example shows the models competency in identifying potential misinformation based on contextual signals and high-risk content.

J Related Work

J.1 Multimodal Rumor Detection

Multimodal rumor detection has significantly advanced since 2014, primarily through the integration of textual, visual, and social context features to enhance detection accuracy. Initial approaches, such asăBoididou et al. (Boididou et al., 2014) and Sabir et al. (Sabir et al., 2018) concentrated on multimedia integrity verification, early methods, like Jin et al. (Jin et al., 2017), used RNN-based models with simple feature concatenation. Later, Tian et al. (Tian et al., 2020) investigated semantic retrieval with autoencoders. MARN was first presented by Wang et al. (Wang and Sui, 2021), who combined BERT and ResNet-18 to improve cross-modal interaction. Chen et al. (Chen et al., 2021)

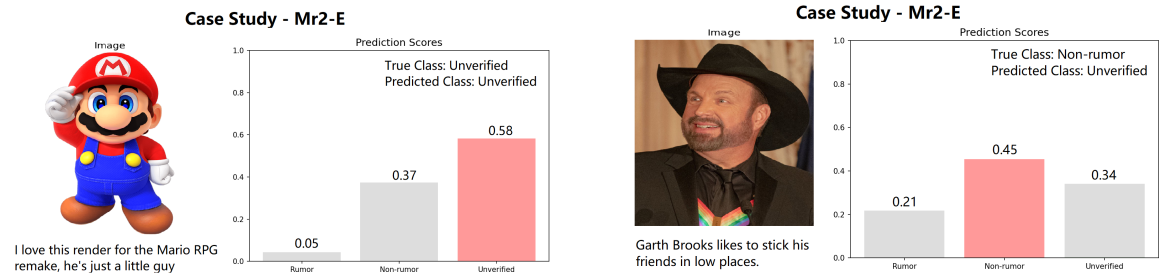

Figure 7: English Case - Unverified Post with Fictional Game Character (Left) and Misclassified Non-rumor Post (Right)

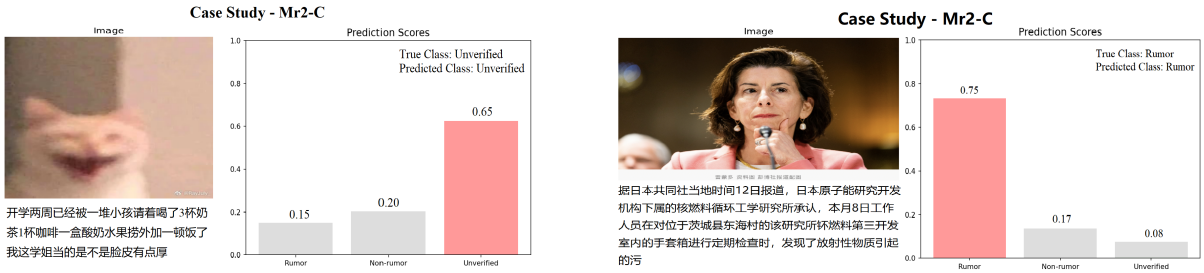

Figure 8: Chinese Case Humorous Unverified Post (Left) and Rumor with News-style Structure (Right)

and Fu et al. (Fu and Sui, 2022) then added self-attention and entity recognition for better fusion. In 2023, models like TDEDA (Han et al., 2023a), CMAC (Zou et al., 2023), CLKD-IMRD (Xu et al., 2023a), MMRDF (Jiang et al., 2023), and CLIP-guided learning (Zhou et al., 2023) introduced adversarial learning, attention mechanisms, and knowledge distillation. Further contributions included MARBERTv2 for Arabic-language rumor detection (Albalawi et al., 2023) and refined attention-based methods from Wang et al. (Wang et al., 2023b). Despite this progress, many models continued to struggle with fine-grained semantic alignment between modalities.

In 2024, several innovations aimed to address these limitations. Guo et al. (Guo et al., 2024) improved cross-modal fusion, while Li et al. (Li et al., 2024b) introduced MKV for semantic mapping and Pang et al. (Pang et al., 2024) explored topic-image alignment. CLIP-based multimodal frameworks such as SARD (Yan et al., 2024) further enhanced semantic integration. Zhou et al. (Zhou et al., 2020a) worked on aligning semantically similar and dissimilar modality pairs. Domain adaptation became a key focus through event-invariant models like EANN (Wang et al., 2018), MKEMN (Zhang et al., 2019), MDDA (Zhang et al., 2021), and MDMN (Zhou et al., 2022), while Ran and Jia (Ran and Jia, 2023) introduced unsupervised cross-domain adaptation using contrastive learning and cross-attention. In parallel, GCNs were adopted by Nanjiang et al. (Nanjiang et al., 2022), Zhong et al. (Zhong et al., 2022), and Sun et al. (Sun et al., 2023b) to combine structural and content-based features, while propagation-based detection gained popularity concurrently. Azri et al. (Azri et al., 2021) used CNN-LSTM architectures. Though most still worked in Euclidean space without hierarchical abstraction or hyperbolic modeling, alignment strategies like SAFE (Zhou et al., 2020b), CARMN (Song et al., 2021), co-attention networks (Wu et al., 2021), and Bi-GRU with image captioning (Wang et al., 2022) addressed semantic inconsistency.

From late 2024 to 2025, attention shifted toward deeper semantic and structural integration. Advances included contrastive alignment (Zhang et al., 2024), deep visual-language fusion networks (DVLFN) (Yang et al., 2023), and knowledge-guided fusion (Sun et al., 2023a). Cross-modality modeling for video content (Li et al., 2024a), stance filtering (Sengan et al., 2024), and contrastive feature learning (Zhou et al., 2025) further pushed the field. Emotion- and transformer-based approaches were explored by Wang et al. (Wang et al., 2023a) and Lv et al. (Lv et al., 2023), while generative and spectral models such as MVAE (Khattar et al., 2019) and FSRU (Lao et al., 2024) introduced new modeling strategies.

In comparison with these work, our proposed RumorCone framework couples hierarchical semantic abstraction and hyperbolic multimodal fusion. Through enabling local fine-grained alignment and global structural coherence, RumorCone corrects semantic disagreement and works stably under low-context and noisy rumor detection scenarios.

J.2 Hierarchical Semantic Learning in Multimodal Models

To better capture the multi-level semantics inherent in multimodal data, several recent works have proposed hierarchical learning mechanisms. Qian et al. (Qian et al., 2021) proposed the Hierarchical Multi-modal Contextual Attention Network (HMCAN) for fake news detection in 2021 by combining multi-modal context with hierarchical textual semantics using BERT and ResNet to process text and images, outperforming state-of-the-art methods. Zhang et al. (Zhang et al., 2023) (2023) proposed the Hierarchical Semantic Enhancement Network (HSEN) for fake news detection, enhancing text and image semantics using hierarchical learning as well as improving inconsistency detection. Ying et al. (Ying et al., 2021) simultaneously incorporated relations of both duplicate and unique modalities and applied multilevel text semantics. Han et al. (Han et al., 2023b) introduced the Cascading Modular Multimodal Cross-Attention Network (CMMCN) in 2023, enabling deep fusion between vision features and text features, overcoming other models with advanced word-level and visual-object level interactions. In 2023, Xu et al. (Xu et al., 2023b) introduced Hierarchically Aggregated Graph Neural Networks (HAGNN), a GNN model aggregating text and propagation structure features, and surpasses baseline models on Weibo and CED datasets. In 2023, Gu et al. (Gu et al., 2023) employed cross-modal co-attention to increase the combination of text and image features to achieve better accuracy on Weibo and Twitter datasets. Despite these advances, most of these approaches rely on Euclidean space, where capturing tree-like or nonlinear relationships between semantic layers remains limited.

J.3 Hyperbolic Embeddings and Geometric Reasoning

Hyperbolic embeddings have been very beneficial in computer vision for encoding hierarchical structures across various domains. Nickel et al. (Nickel and Kiela, 2017) introduced Poincaré embeddings that compress symbolic hierarchies more effectively compared to their Euclidean equivalents. Following that, Hao et al. (Guo et al., 2021) extended hyperbolic embeddings to multi-modal knowledge graphs with the help of hyperbolic graph convolutions for enhanced entity alignment. Hi-Mapper by Kwon et al. (Kwon et al., 2024) enhances scene understanding via hierarchies of visual information learned in hyperbolic space, whereas Kong et al.'s HyperLearner (Kong et al., 2024) blends synthetic captions with hyperbolic learning for open-world object detection boost. Similarly, Kim et al.'s HYPE method (Kim et al., 2024) uses hyperbolic entailment to filter out noisy image-text pairs to boost self-supervised learning.

Other notable contributions include Sinha et al.'s HypStructure (Sinha et al., 2024), which injects label hierarchies into learned features via hyperbolic regularization to generalize better. Dhingra et al. (Dhingra et al., 2018) applied hyperbolic embeddings to unsupervised learning of text representation, whereas Guo et al. (Guo et al., 2022) addressed training problems in hyperbolic neural networks by clipping Euclidean features, enhancing robustness in classification. Khrulkov et al. (Khrulkov et al., 2020) demonstrated dominance of hyperbolic embeddings over Euclidean and spherical embeddings in computer vision. Liu et al. (Liu et al., 2020) proposed a hyperbolic visual embedding network with better zero-shot recognition via more compact embeddings. Desai et al. (Desai et al., 2023) introduced MERU, a vision-language model to learn hierarchical entailment relationships in Lorentzian space, a hyperbolic manifold extension.

Previous work mostly aims at vision-language tasks or knowledge graphs, our **RumorCone** framework is more general as it allows hierarchical hyperbolic modeling to enable cross-modal misinformation detection with a novel fusion mechanism.