Beyond Single-View Detection: A Dual-Space Reasoning Framework for Interpretable Harmful Meme Understanding

Wenqing Hou, Hongkui Tu, Ye Wang, Yue Zhang,
Yuying Liu, Dong Zhu, Liqun Gao*, **Bin Zhou**
College of Computer Science and Technology, National University of Defense Technology
{houwenqing, gaoliqun15}@nudt.edu.cn
https://github.com/icvplayer/BPDMoE-Hate

Abstract

The identification of harmful memes extends beyond a mere classification task, encompassing challenges related to multi-perspective semantic comprehension and hierarchical reasoning. Prevailing approaches predominantly depend on modal alignment or black-box classifiers, which fail to capture implicit biases and lack interpretability. In this study, we propose BPDMoE-Hate, a novel framework grounded in dual-space mixture-of-experts, which innovatively conceptualizes harmful meme detection as an integrated process of "viewpoint decoupling and hierarchical fusion". Our approach generates adversarial binary perspectives via Visual-Language Models (VLMs) and incorporates an adaptive viewpoint gating to facilitate viewpoint selection, thereby enabling the model to autonomously discern implicit semantic inclinations. Moreover, we propose the Hyperbolic-Euclidean space expert to effectively capture the hierarchical structural relationships and semantic correlations between multimodal and viewpoint features, thereby enabling interpretable reasoning at the geometric representation level. Empirical evaluations conducted on three mainstream datasets demonstrate that BPDMoE-Hate not only substantially surpasses existing methodologies in performance but also offers visual explanations for viewpoint selection and hierarchical structuring, thereby advancing the field of interpretable multimodal content analysis.

1 Introduction

The advancement of social media platforms has enhanced the capacity for individuals to express their emotions (Hermida and Santos, 2023; Liu et al., 2024a); however, it has concurrently contributed to the propagation of detrimental information. A notable example includes internet memes, which have gained widespread circulation in recent years.

*Corresponding author

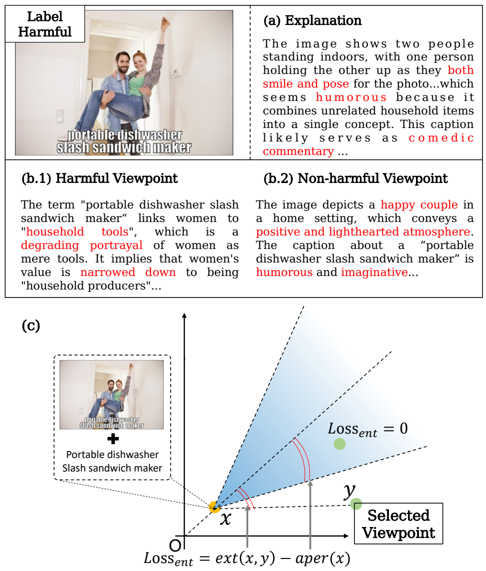


Figure 1: Binary viewpoints and entailment loss.

Typically, a meme comprises an image paired with concise textual content conveying a particular viewpoint. Certain users exploit this format to disseminate hate speech, thereby facilitating and intensifying violent (Mei et al., 2024) conduct in real-world contexts.

Viewpoint decoupling. Previous hate speech meme detection methods either adopt a single classifier (Hebert et al., 2024; Lu et al., 2024) or cross-modal alignment (Yang et al., 2024) to narrow the inter-modality semantic gap, but these black-box models (Lin et al., 2024) exhibit limited interpretability. Recently, studies have leveraged VLMs with zero-shot prompting (Kojima et al., 2022) and multi-agent debate frameworks (Park et al., 2024; Zheng et al., 2024) to interpret harmful memes (Cao et al., 2023; Ji et al., 2024; Lin et al., 2025), and even incorporated explanatory text (Hee and Lee, 2025) or debate outcomes (Lin et al., 2024; Zhou et al., 2025) into model training for performance improvement. However, these interpretative approaches still cannot model multi-perspective se-

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 10526–10544
July 2-7, 2026 ©2026 Association for Computational Linguistics

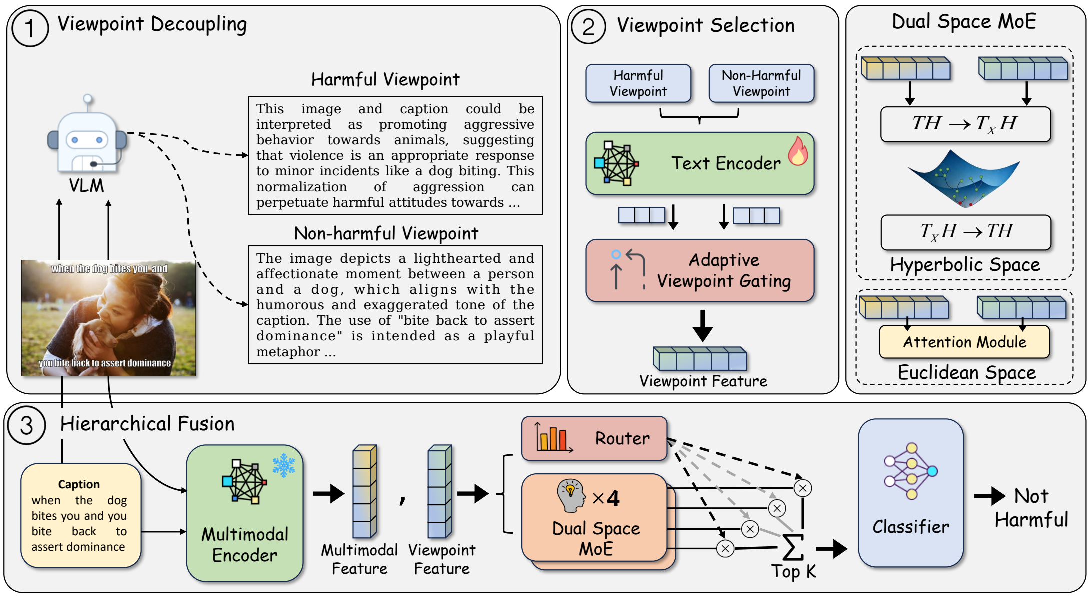


Figure 2: Our overall framework: generate decoupled binary viewpoints via the VLM, select viewpoints using AVG, and then leverage DSMoE to learn dual-space features for classifier-based prediction. The dashed box delineates the mapping and inverse mapping operations that are unique to hyperbolic space.

mantic conflicts, and are thus prone to introducing the model's subjective bias towards memes.

As illustrated in Figure 1, the text within the meme diminishes the value of women by portraying them as mere "household tools". Figure 1 (a) incorporates the model's own subjective judgment, erroneously interpreting the content as humorous. Consequently, integrating such explanations into the training process risks propagating subjective biases, leading to misclassification. Panels (b.1) and (b.2) represent the explanatory approach we propose, which involves a more in-depth analysis of memes from two distinct perspectives. This decoupled adversarial viewpoint partially mitigates the model's inherent bias. Nonetheless, only one of these perspectives accurately reflects the true nature of the harmful meme. Consequently, this raises an important question: how can the model be trained to autonomously select the appropriate viewpoint?

Hierarchical reasoning. Beyond the viewpoint selection challenge, a more critical issue unaddressed by existing methods is the lack of a systematic framework to model the hierarchical semantic relationships between multimodal content and interpretative perspectives. Regarding the meme depicted in Figure 1, from the perspective of cognitive logic (Van Ditmarsch et al., 2008), human observers first synthesize information from both components (image and title) to achieve comprehension, subsequently generating clear and specific views on it. These perspectives represent a more profound understanding of the multimodal information—comprising both images and texts—thereby establishing a hierarchical structure (Vendrov et al., 2015) that progresses from "multimodal content" to "human viewpoints". Given that hyperbolic space (Desai et al., 2023; Pal et al., 2024) near the origin encodes more general information, while regions closer to the boundary convey more specific attributes, it is appropriate to represent multimodal content and viewpoints as "roots" situated near the origin and "leaves" positioned closer to the periphery within the hyperbolic space.

This study revisits the problem of harmful meme detection from the perspective of geometric representation learning. We hypothesize that understanding harmful memes entails a reasoning process of "viewpoint decoupling-hierarchical fusion". Specifically, we propose BPDMoE-Hate, a **B**inary **P**erspectives **D**ual-space **M**ixture-**o**f-**E**xperts framework comprising two core components: an Adaptive Viewpoint Gating (AVG) module for viewpoint selection and a Dual-Space MoE (DSMoE) module for hierarchical feature fusion. In the viewpoint decoupling phase, dual viewpoints are generated via adversarial prompting, and the AVG module is designed to enable adaptive viewpoint selection, thereby empowering the model to discern semantic veracity. Subsequently, in the hierarchical fusion phase, the hierarchical structure of multimodal and viewpoint features is modeled in hyperbolic space, while semantic relationships are captured in Euclidean space. Finally, the DSMoE module dynamically integrates these two types of representations to form a unified and discriminative feature space.

Our contributions are as follows: 1) We propose a multi-view decoupled reasoning framework, which formalizes the task as a process of view generation, selection and fusion. 2) We design a DSMoE to provide an interpretable geometric foundation for multimodal reasoning. 3) Experimental results verify the superior performance of our approach.

2 Preliminaries

In this section, we provide a concise overview of the essential concepts underlying hyperbolic geometry; more details can be found in (Chami et al., 2019). Hyperbolic space is characterized as the unique complete, simply connected Riemannian manifold that is isotropic and exhibits constant negative sectional curvature (Wang et al., 2024). The curvature parameter quantifies the extent to which hyperbolic space diverges from Euclidean flatness. The models used to describe hyperbolic space mainly include the Lorentz model and the Poincaré model. This work focuses exclusively on the Poincaré model.

Consider the n-dimensional Poincaré sphere endowed with a constant negative curvature $-c(c > 0)$. The associated Riemannian manifold can be characterized as $\mathcal{H}^{n,c} = \{x \in \mathcal{H}^n \mid \|x\|^2 < \frac{1}{c}\}$. For any point $x_{\mathcal{H}} \in \mathcal{H}^{n,c}$ within this hyperbolic space, there exists a corresponding tangent space $\mathcal{T}_x\mathcal{H}^{n,c}$, which serves as a local, first-order approximation of the manifold at $x_{\mathcal{H}}$. The hyperbolic space and its tangent space at $x_{\mathcal{H}}$ are related through the following mapping:

$$exp_x^c(v) = v \oplus_c (tanh(\sqrt{c}\frac{\lambda_x^c \|v\|}{2})\frac{v}{\sqrt{c}\|v\|}) \quad (1)$$

$$log_x^c(y) = \frac{2tanh^{-1}(\sqrt{c}\|-x\oplus_c y\|)(-x\oplus_c y)}{\sqrt{c}\lambda_x^c \|-x\oplus_c y\|} \quad (2)$$

In this context, $v \in \mathcal{T}_x\mathcal{H}^{n,c}$, $x, y \in \mathcal{H}^{n,c}$. $\lambda_x^c = \frac{2}{1-c\|x\|^2}$ is the conformal factor, $\oplus_c$ denotes Möbius addition (Ungar, 2001). The exponential operation maps $\mathcal{T}_x\mathcal{H}^{n,c}$ to $\mathcal{H}^{n,c}$, while the logarithmic operation maps $\mathcal{H}^{n,c}$ to $\mathcal{T}_x\mathcal{H}^{n,c}$. The distance within the Poincaré sphere is defined as the shortest path between two points x and y:

$$d_{xy}^c = \frac{2}{\sqrt{c}}\mathrm{arc}tanh(\sqrt{c}\|-x\oplus_c y\|) \quad (3)$$

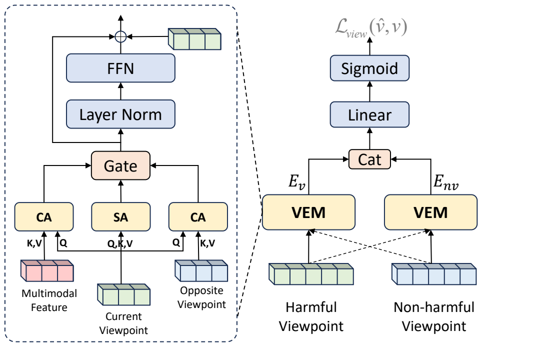


Figure 3: Adaptive viewpoint gating. VEM represents our viewpoint enhancement module, CA signifies Cross-Attention, while SA denotes Self-Attention.

3 Method

3.1 Viewpoint Decoupling

The overall architecture is illustrated in Figure 2. To mitigate the subjective biases (Ye et al., 2024) present in models during the generation of VLM explanations, we conceptualize the generation of viewpoints as an adversarial semantic completion task. Concretely, we design a pair of adversarial prompt templates that direct the VLM to produce two opposing textual descriptions—one harmful and one harmless—for the identical meme, as illustrated in step 1 of Figure 2. This approach can be interpreted as a semantic decoupling of the original multimodal content, with the objective of disentangling implicit semantic dimensions that may contribute to ambiguity in the final evaluation. The corresponding prompt templates are presented in Appendix B.

3.2 Viewpoint Selection

We use the text encoder to encode the decoupled binary viewpoints. Nevertheless, the model must identify the viewpoint that more accurately corresponds to the true semantics between the two contrasting perspectives, which is the prerequisite for achieving reliable hierarchical fusion.

3.2.1 Adaptive Viewpoint Gating

Inspired by (Zhao et al., 2022), we design an AVG to conduct the selection of viewpoint features. As illustrated in Figure 3, the two viewpoint feature vectors produced by the text encoder are denoted as $H_v, H_{nv} \in \mathcal{R}^{B\times L\times N}$. Where B represents the batch size, L denotes the length of the text sequence, N represents the output dimension of the hidden layer. These two vectors are individually enhanced through the core module VEM (Viewpoint Enhancement Module), allowing each viewpoint feature to interactively reference its complementary viewpoint as well as the original multimodal context. Consequently, this process facilitates the acquisition of more discriminative representations.

Subsequently, the enhanced feature representations E_v and E_{nv} are derived and subsequently fed into the classification layer to produce the final viewpoint selection weights $W_v \in \mathcal{R}^{B\times 1}$. This process is formally represented by the following equation:

$$W_v = Sigmoid(Linear(Cat(E_v, E_{nv}))) \quad (4)$$

Here, "$Cat(\cdot)$" represents feature concatenation. After obtaining the weights for the viewpoint selection, for each sample ${h^i}_v \in \mathcal{R}^{L\times N}$, ${h^i}_{nv} \in \mathcal{R}^{L\times N}$ and ${w^i}_v \in \mathcal{R}^{1\times 1}$, where $i \in B$, we perform the viewpoint selection through the following formula:

$$ {h^i}_s = {w^i}_v \cdot {h^i}_v + (1 - {w^i}_v) \cdot {h^i}_{nv} \quad (5)$$

Moreover, we develop a viewpoint selection loss, wherein the true labels are consistent with the downstream harmful detection task, forcing the selection mechanism to be congruent with the final task objective:

$$\mathcal{L}_{view} = \mathcal{L}_{CE}(\hat{v}, v) \quad (6)$$

Where $\mathcal{L}_{CE}$ denotes the cross-entropy loss, $\hat{v}$ signifies the prediction generated by AVG for the selection of viewpoints, and v corresponds to the true label.

3.2.2 Viewpoint Enhancement Module

Suppose the target sample currently input is ${h^i}_v$, that is "Current Viewpoint", then ${h^i}_{nv}$ is "Opposite Viewpoint", while the features from the multimodal encoder are $m^i \in \mathcal{R}^{M\times N}$. Firstly, perform multi-head self-attention and cross-attention operations on the current features, that is, the SA and CA modules in Figure 3. Here, we use CA as an example. The Queries are derived from ${h^i}_v$, represented as $Q_t = I_t W_q$, while the Keys and Values come from m^i, represented as $K_s = I_s W_k$ and $V_s = I_s W_v$, where $W_q, W_k, W_v \in \mathcal{R}^{N\times N}$ are learnable parameters. The CA can then be formulated as:

$$\begin{aligned} o^i_{cm} &= \text{Softmax}\left(\frac{Q_t K_s^T}{\sqrt{d_k}} V_s\right) \\ &= \text{Softmax}\left(\frac{I_t W_q W_k^T I_s^T}{\sqrt{d_k}}\right) I_s W_v \end{aligned} \quad (7)$$

In the SA module, the Queries, Keys, and Values are all derived from the same feature. We represent the vectors produced by the SA and an additional CA as o^i_c and o^i_{co}, respectively. These are then combined using a "Gate" followed by a feed-forward network (FFN) to achieve an enhanced representation $e^i_v \subset E_v$ (Appendix E.8 demonstrates the effectiveness of VEM):

$$\begin{cases} e^i_v = h^i_v + e^i + FFN(LN(e^i)) \\ e^i = w^i_e \cdot o^i_c + (1 - w^i_e) \cdot (o^i_{cm} + o^i_{co})/2 \\ w^i_e = Sigmoid(Linear(Cat(o^i_c, o^i_{cm}))) \end{cases} \quad (8)$$

3.3 Dual Space MoE

3.3.1 Hyperbolic Space Experts

To construct the hierarchical structure within hyperbolic space, we initially employ Eq.1 to project the selected viewpoint and multimodal features onto the hyperbolic manifold. As indicated in (Wang et al., 2024), it is necessary to first apply a linear transformation to the features to reduce their dimensionality, thereby enhancing the representational capacity of the hyperbolic manifold. Given a weight matrix $W \in \mathcal{R}^{m\times n}$ and a bias term $b \in \mathcal{R}^{1\times 1}$ associated with a linear layer, the matrix multiplication operation within the Poincaré ball is performed via Möbius multiplication:

$$W \oplus_c x = \exp_0^c(W \cdot \log_0^c(x)) \quad (9)$$

In the context of biased addition, the vector b represents a vector located in the tangent space $\mathcal{T}_0\mathcal{H}^{n,c}$. It is necessary to transport this vector to the tangent space $\mathcal{T}_x\mathcal{H}^{n,c}$ at the point x, after which it is projected onto the hyperbolic manifold:

$$T^c_{0\to x}(b) = \log_x^c(x \oplus_c \exp_0^c(b)) \quad (10)$$

Where $x^{\mathcal{H}} \oplus_c b = \exp_x^c(T^c_{0\to x}(b)) = \frac{\lambda_0^c}{\lambda_x^c} b$. As illustrated in Figure 1(c), we further adopt the entailment loss (Desai et al., 2023) to constrain viewpoint features within the entailment cone centered on multimodal features, thus geometrically encoding the "multimodal entailment viewpoint" reasoning relationship explicitly. Let x and y represent the multimodal and viewpoint vectors embedded in hyperbolic space, respectively. Here, x should lie closer to the hyperbolic origin, while y should reside within the entailment cone originating at x. The half-aperture angle $\mathrm{aper}(x)$ of x is defined as follows:

$$aper(x) = \arcsin\left(\frac{r_{\min}\sqrt{c}}{\tanh(\frac{\sqrt{c}d_{ox}^c}{2} + \varepsilon)}\right) \tag{11}$$

Here $r_{\min} = 0.1$ denotes the minimum radius of the containment cone, $\varepsilon = 1\times10^{-8}$, c is a learnable curvature value, initialized to 1.0 and decreasing during training. The external angle $ext(x, y) = \pi - \angle Oxy$, thus $\angle Oxy$ can be obtained by the following formula:

$$\angle Oxy = \cos^{-1}\left(\frac{\cosh(D_{ox}^c)\cosh(D_{xy}^c) - \cosh(D_{oy}^c)}{\sinh(D_{ox}^c)\sinh(D_{xy}^c) + \varepsilon}\right) \tag{12}$$

Where $D_{AB}^c = \sqrt{c}\, d_{AB}^c$, while d_{AB}^c corresponds to the hyperbolic distance between points A and B, as determined by Eq.3. Then, the final entailment loss is expressed as:

$$\mathcal{L}_{ent} = \max(0, ext(x, y) - aper(x)) \tag{13}$$

Moreover, we implement CA and SA for hyperbolic features via a "project-infer-project" paradigm as shown in Figure 2: hyperbolic features are first mapped back to Euclidean space for attention computations, then reprojected to hyperbolic space; the final fused hyperbolic features are mapped to Euclidean space as described by Eq.2.

3.3.2 Euclidean Space Expert

Euclidean space experts are responsible for modeling the semantic connections between multimodal and viewpoints. In this instance, the projection and back-projection operations are excluded, and only the attention mechanism is retained.

3.3.3 Hierarchical Fusion and Prediction

Our DSMoE innovatively integrates experts from different spaces into the MoE (Liu et al., 2025) layer, which is composed of a Router and n expert networks, where we set $n = 4$. The experts with odd indices operate in Euclidean space, while those with even indices function in hyperbolic space. As shown in Figure 2, the Router (a standard transformer block), dynamically integrates experts from two distinct domains, enabling the model to adaptively integrate hierarchical reasoning and semantic correlation information, thereby facilitating hierarchical fusion subsequent to viewpoint decoupling. Subsequently, a classification layer is employed to derive the probabilities of n experts being assigned and top 2 experts are chosen to participate in the computation. The detailed formula is as follows:

$$O_{exp} = \sum_{i=1}^{n} w_{\exp}^i \cdot o_{\exp}^i \tag{14}$$

Where $o_{\exp}^i$ denotes the output from the ith expert. Ultimately, the harmful meme prediction is produced by the classification layer. We employ the cross-entropy loss between the model's predicted category and the actual category, expressed as $\mathcal{L}_{task} = \mathcal{L}_{CE}(\hat{y}, y)$.

In addition, we also introduce contrastive learning loss $\mathcal{L}_{InfoICE}$ (Oord et al., 2018) to better distinguish the positive and negative samples in the sampling process, and load balancing loss $\mathcal{L}_{balance}$ (Fedus et al., 2022) to promote the balanced loading of experts. For detailed information, please refer to Appendix C and E.4. The overall training loss is as follows:

$$\begin{aligned}\mathcal{L}_{all} = \mathcal{L}_{task} + \alpha\mathcal{L}_{view} + \beta\mathcal{L}_{ent} + \\ \gamma\mathcal{L}_{balance} + \eta\mathcal{L}_{InfoICE}\end{aligned} \tag{15}$$

4 Experiments

4.1 Evaluation Dataset and Baselines

We evaluate BPDMoE-Hate on three widely used hate meme datasets, including Facebook's Hateful Meme (FHM) (Kiela et al., 2020), Multimedia Automatic Misogyny Identification (MAMI) (Fersini et al., 2022), and Harmful Meme (HarMeme) (Pramanick et al., 2021). Furthermore, we demonstrate the effectiveness of our method by comparing it with some state-of-the-art meme detection models (separated by a double line in Table 1), which include: 1) VLMs 2) pure text classification models and multimodal classification models 3) harmful meme detection framework. For detailed supplementary information on the datasets and references to the models, please refer to Appendix D.

4.2 Implementation Details

To guarantee the reliability of the decoupling of viewpoints, we employ Qwen2.5-VL-32B-Instruct

Model	FHM		MAMI		HarMeme	
	AUC	ACC	AUC	ACC	AUC	ACC
Llama-3.2-V	-	63.38 $\pm$0.83	-	69.14 $\pm$0.69	-	69.21 $\pm$1.61
Llava-1.5	-	52.34 $\pm$0.81	-	53.72 $\pm$1.45	-	55.37 $\pm$1.03
Qwen2.5-VL	-	73.52 $\pm$0.16	-	69.78 $\pm$0.25	-	63.33 $\pm$0.13
BERT-base	64.98 $\pm$0.61	57.86 $\pm$0.68	71.56 $\pm$0.65	64.28 $\pm$0.52	81.38 $\pm$0.88	75.31 $\pm$1.19
RoBERTa-large	64.40 $\pm$1.29	58.56 $\pm$0.43	72.46 $\pm$0.65	66.08 $\pm$1.15	81.72 $\pm$0.83	76.38 $\pm$0.90
FLAVA-full	78.60 $\pm$0.74	70.02 $\pm$1.39	81.24 $\pm$0.55	70.10 $\pm$0.61	85.45 $\pm$1.06	79.72 $\pm$1.78
VisualBERT*	68.71 $\pm$1.02	61.48 $\pm$1.19	78.71 $\pm$0.59	71.06 $\pm$0.94	80.46 $\pm$1.04	75.31 $\pm$1.44
ViLBERT*	73.05 $\pm$0.62	64.70 $\pm$1.12	77.71 $\pm$1.20	69.48 $\pm$1.00	84.11 $\pm$0.88	78.70 $\pm$1.17
BLIP2	63.52 $\pm$0.62	58.18 $\pm$0.96	82.05 $\pm$0.20	65.50 $\pm$3.53	89.94 $\pm$0.14	80.62 $\pm$1.84
ALBEF*	79.40 $\pm$0.53	70.58 $\pm$0.50	83.24 $\pm$0.93	72.77 $\pm$1.00	85.49 $\pm$1.23	80.99 $\pm$0.80
Mod-HATE	64.50 $\pm$0.19	58.00 $\pm$1.07	67.40 $\pm$0.46	61.00 $\pm$2.22	73.40 $\pm$0.27	69.40 $\pm$0.42
PromptHate	76.76 $\pm$0.95	67.82 $\pm$1.23	76.21 $\pm$1.05	68.08 $\pm$0.58	87.51 $\pm$0.74	79.38 $\pm$1.72
Pro-Cap	80.87 $\pm$0.66	72.28 $\pm$0.90	82.53 $\pm$0.49	73.06 $\pm$0.82	90.25 $\pm$0.54	83.25 $\pm$1.00
ExplainHM	82.32 $\pm$1.12	72.22 $\pm$1.62	79.07 $\pm$2.13	71.03 $\pm$0.88	75.58 $\pm$4.92	77.16 $\pm$1.98
IntMeme	81.50 $\pm$1.11	71.52 $\pm$1.49	81.89 $\pm$1.15	72.30 $\pm$1.79	89.35 $\pm$1.22	81.92 $\pm$2.47
BPDMoE-Hate	**83.71** $\pm$0.39	**75.18** $\pm$1.38	**87.84** $\pm$0.54	**76.70** $\pm$0.80	**94.11** $\pm$0.28	**86.10** $\pm$0.94

Table 1: The results are presented as the "mean $\pm$ standard deviation" of the outcomes derived from five distinct random seeds. * indicates the results are from (Cao et al., 2023). We use the same VLM to generate the explanations.

(Qwen2.5-VL) as the viewpoint generation model. The text encoder and multimodal encoder are instantiated using RoBERTa-large and BLIP2, respectively. During training, the multimodal encoder parameters are frozen, while the final two layers of the text encoder are fine-tuned over 4 epochs. The loss function incorporates weighted components with coefficients $\alpha, \beta, \gamma, \eta$ set to 0.5, 1×10^{-2}, 1×10^{-3}, and 0.1, respectively. A learning rate of 1×10^{-5} is adopted, and the entire training procedure is conducted for 10 epochs (batch size is set to 48) utilizing one NVIDIA A6000 GPU. Model performance is evaluated using Accuracy (ACC) and the Area Under the Receiver Operating Characteristic Curve (AUC) as metrics.

4.3 Main Results

As presented in Table 1, the BPDMoE-Hate model demonstrated superior performance across all three datasets. Notably, while explanation-based approaches such as ExplainHM and IntMeme yielded competitive results—particularly with Pro-Cap attaining an AUC of 90.25 on the HarMeme dataset—our method, by effectively integrating two distinct perspectives, achieved a higher AUC of 94.11. This enhancement substantiates the efficacy of the dual-view decoupling and selection mechanism in identifying implicit harmful content. Furthermore, the consistent outperformance on the MAMI dataset, which focuses on explicitly hateful content directed towards women, as well as the FHM dataset, characterized by mixed content types, suggests that the proposed dual-space hierarchical fusion framework possesses strong generalizability across diverse tasks.

4.4 Ablation Study

As shown in Figure 4, we demonstrate the contributions of different modules. The simultaneous removal of AVG and DSMoE resulted in a marked decline in performance, indicating that AVG plays a critical role in selection following perspective decoupling, whereas DSMoE is essential for hierarchical fusion. Both components are therefore indispensable. This highlights the importance of the two-stage architecture of our framework, termed "decoupling-fusion". The elimination of AVG alone led to a more pronounced performance degradation, underscoring its critical role in mitigating model biases and autonomously identifying the accurate perspective. Interestingly, the removal of DSMoE alone on the MAMI dataset resulted in a slight increase in AUC. We believe that the misogynistic content within this dataset is predominantly expressed explicitly, thereby reducing the reliance on complex hierarchical reasoning.

4.5 The Significance of Viewpoints

Given the importance of uncovering the latent information within memes for enhancing model decision-making, we conduct a series of experiments to assess the influence of various interpreta-

Viewpoint Type	FHM		MAMI		HarMeme	
	AUC	ACC	AUC	ACC	AUC	ACC
BPDMoE-Hate	83.71 $\pm$0.39	75.18 $\pm$1.38	87.84 $\pm$0.54	76.70 $\pm$0.80	94.11 $\pm$0.28	86.10 $\pm$0.94
Random Select	79.78 $\pm$1.21	69.88 $\pm$1.43	86.54 $\pm$0.87	73.04 $\pm$1.29	89.95 $\pm$0.90	83.33 $\pm$1.47
Harmful View	82.06 $\pm$0.56	72.78 $\pm$1.88	87.12 $\pm$0.50	74.04 $\pm$1.31	90.29 $\pm$1.07	83.79 $\pm$1.11
Non-harmful View	79.35 $\pm$0.55	69.04 $\pm$1.09	86.60 $\pm$1.05	73.90 $\pm$1.71	89.98 $\pm$1.45	83.45 $\pm$1.33
No View	70.16 $\pm$2.05	61.98 $\pm$1.17	82.49 $\pm$1.14	67.74 $\pm$1.22	92.74 $\pm$0.11	84.58 $\pm$1.89
Only Explanation	82.80 $\pm$0.16	73.18 $\pm$1.81	87.10 $\pm$0.61	75.38 $\pm$1.09	92.75 $\pm$0.24	85.76 $\pm$0.96

Table 2: The influence of different viewpoints. We set up 5 types of viewpoints for training, including: 1) Randomly select one of the viewpoints. 2) Only use harmful viewpoints. 3) Only use harmless viewpoints. 4) Remove both types of viewpoints. 5) Replace both types of viewpoints with explanations for memes.

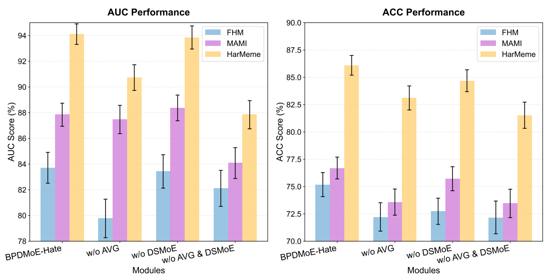


Figure 4: The impact of different modules on the overall performance of the model. We gradually removed AVG and DSMoE.

tions. The results, presented in Table 2, indicate that employing a single viewpoint leads to a noticeable decline in performance. Furthermore, the approach of randomly selecting viewpoints, which disrupts the AVG's selection strategy, yields even poorer outcomes compared to using only one viewpoint. The performance without using any viewpoint is the worst, underscoring the critical role of hidden information extraction in meme detection. Conversely, utilizing solely model explanations produces comparatively better results. These findings demonstrate that the concurrent integration of both viewpoints exerts a more beneficial impact on the model's decision-making process.

4.6 The Significance of DSMoE

In this section, we conducted tests using the same random seed, and the experimental results are presented in Figure 5(a). It is evident that the AUC and ACC values for both "Only HS" and "Only ES" are inferior to those of BPDMoE-Hate, indicating that our model effectively captures complementary information from different spaces. The absence of either spatial feature representation adversely affects the model's performance. Furthermore, the overall performance of "Only HS" surpasses"Only ES", suggesting that hierarchical structure model-

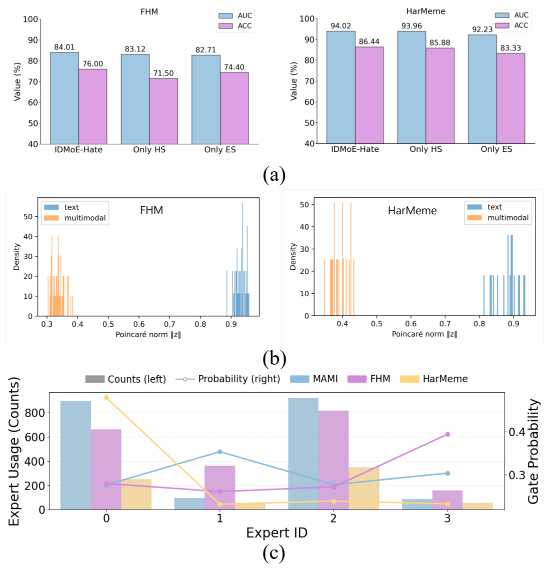
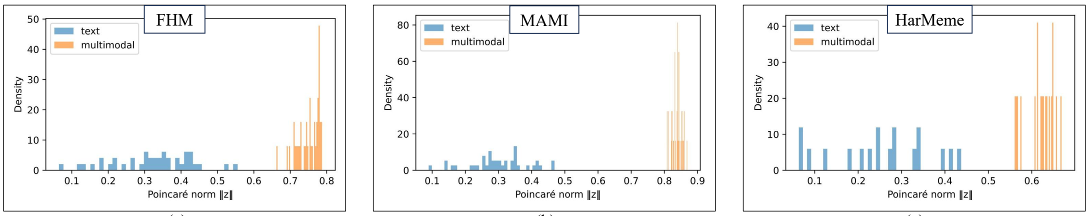


Figure 5: The importance of DSMoE. In (a), "Only ES" indicates replacing all experts with Euclidean experts, while "Only HS" represents "only hyperbolic space experts". (b) represents the hierarchical structure in hyperbolic space. "text" represents the selected viewpoint feature. (c) denotes the frequency of activation for each expert alongside the corresponding mean probability.

ing in hyperbolic space contributes more positively to the model's gains.

To validate the geometric assumptions underpinning the hierarchical structure modeling in hyperbolic space, we utilize the visualization method from (Pal et al., 2024) and present the learned hyperbolic space structure through low-dimensional visualization (showing the spatial norm distribution of the test set samples in the form of histograms). As shown in Figure 5(b), multimodal features consistently cluster near the origin of the hyperbolic space, while viewpoint features are located farther away. This notable geometric distinction offers quantitative evidence that the hierarchical structure of the "multimodal entailment viewpoint features"



(a) (b) (c)

Method	AUC (↓2.59)	ACC (↓2.04)
BPDMoE-Hate	83.71±0.39	75.18±1.38
Reverse Hierarchy	81.12±1.25	73.14±0.73

Method	AUC (↑0.75)	ACC (↓2.28)
BPDMoE-Hate	87.84±0.54	76.70±0.80
Reverse Hierarchy	88.59±0.43	74.42±2.18

Method	AUC (↓0.54)	ACC (↓0.34)
BPDMoE-Hate	94.11±0.28	86.10±0.94
Reverse Hierarchy	93.57±0.42	85.76±0.81

Figure 6: The influence of the reverse hierarchical structure. We forcibly reversed the hierarchical relationship between the viewpoint ("text") and the multimodal information.

has been effectively represented.

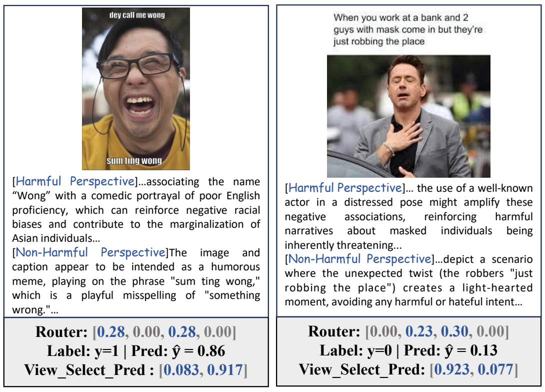
As shown in Figure 5(c), experts in the hyperbolic space are more frequently activated on most datasets. This finding suggests that the model predominantly depends on hyperbolic space representations, which are capable of encoding hierarchical semantic relationships, during decision-making processes. From the perspective of the activation probabilities of experts, except for the HarMeme dataset, which tends to assign greater weight to the hyperbolic space expert with ID 0, the other two datasets exhibit a preference for allocating higher probabilities to the Euclidean space experts. This observation clearly demonstrates the flexibility of our framework in selecting different experts.

4.7 Reverse Hierarchical Structure

We performed a causal intervention experiment by repositioning the viewpoint features to the origin of the hyperbolic space while situating the multimodal features at the periphery. The outcomes, presented in Figure 6, demonstrate a marked decrease in performance across the three datasets. These findings indicate that the observed positive hierarchical relationship is not merely a coincidental correlation within the data but constitutes a causal factor contributing to the model's superior performance. Furthermore, we also verify the effect of removing this hierarchical relationship, detailed information is provided in Appendix E.6.

4.8 Cross-Domain and Cross-Cultural Generalization

We evaluate the cross-domain performance by training the model on the FHM dataset (general harmful meme detection) and directly testing on the MAMI dataset (misogyny-specific meme detection), with results reported in Table 3. Despite a natural performance drop due to the domain shift from general



(a) (b)

Figure 7: Case illustration. "Router" represents the activated experts and their assigned probabilities, while "View_Select_Pred" indicates the probability of each viewpoint being selected.

Method	MAMI	
	ACC	AUC
IntMeme	57.20	64.00
BPDMoE-Hate	**62.50**	**72.37**

Table 3: Cross-domain transfer performance (trained on FHM, tested on MAMI), here we set the random seed to 42.

memes to misogyny-specific content, BPDMoE-Hate maintains a substantial advantage over the baseline IntMeme, with absolute gains of 5.30% in ACC and 8.37% in AUC. This result verifies the robust cross-domain generalization of our method.

To test the cross-cultural adaptability, we further evaluate BPDMoE-Hate on ToxiCN (Lu et al., 2024), a Chinese harmful meme dataset, using VLM-generated Chinese decoupled viewpoints as input. The comparison with state-of-the-art baselines is shown in Table 4. The results demonstrate that our method achieves significant performance

Method	ToxiCN	
	ACC	F1
PromptHate	76.04	67.45
Pro-Cap	75.70	71.36
ExplainHM	75.20	67.60
ALARM (GPT-4o)	77.45	67.87
BPDMoE-Hate	**82.75**	**71.29**

Table 4: Cross-cultural performance on ToxiCN Chinese harmful meme dataset (Set the random seed to 42).

advantages on Chinese memes, outperforming all compared baselines by a clear margin. Specifically, BPDMoE-Hate obtains the highest ACC of 82.75%, with an absolute improvement of 5.30% over the strongest baseline ALARM (Lang et al., 2025). Collectively, both experiments confirm the strong generalization ability of BPDMoE-Hate across different domains and cultural contexts.

4.9 Case Study

Figure 7(a) is the malicious pun on the Asian surname “Wong” by deliberately spelling “something wrong” as “sum ting wong” to imitate a stereotypical Asian accent is offensive and defamatory to the Asian community. Our harmful viewpoint correctly explains this point. Additionally, our Router utilizes hyperbolic space experts to accurately identify such harmful memes, and AVG successfully select the correct viewpoint. Figure 7(b) depicts the actor portraying Iron Man exhibiting relief upon confronting the masked criminal. Since typical robberies do not necessitate superhero intervention, this scenario conveys a humorous and teasing tone. These observations further substantiate the efficacy of our proposed framework.

5 Conclusion

This paper presents BPDMoE-Hate, a dual-space viewpoint decoupled reasoning framework for detecting harmful memes. Our framework utilizes the generated dual viewpoints as the input for semantic decoupling, and employs an AVG to autonomously determine the semantic authenticity, effectively alleviating the model’s subjective bias. Furthermore, we designed a dual-space MoE, which explicitly models the hierarchical entailment relationship between multimodal and perspective features in the hyperbolic space, and learns semantic associations in the Euclidean space, achieving the synergy of structured reasoning and semantic matching.

Acknowledgements

This work is supported by National Key Research and Development Plan in China (2023YFC3306100)

Limitations

The limitations of this study are as follows. Our proposed framework depends on a robust VLM to produce high-quality binary perspectives. Although the implementation of a perspective selection mechanism mitigates bias to some extent, the diversity and comprehensiveness of the generated perspectives remain constrained by the inherent cognitive limitations of the VLM. In Appendix E.3, we provide a detailed argument supporting the decoupling perspective generated by the small model. Future research will investigate retrieval-augmented generation (RAG) techniques to enhance the quality and controllability of perspective generation in smaller models.

Ethical Considerations

Data Privacy and Compliance: The FHM, MAMI and HarMeme datasets were used in strict accordance with the original authors’ terms of use. No additional personal data was collected, and all multimodal content was anonymized to protect user privacy. All human evaluation procedures in this study were conducted in strict compliance with ethical guidelines for human subject research. **Expected Use and Misuse Prevention:** This framework is exclusively for harmful meme detection to build a safer online environment. Unethical/illegal use is strictly prohibited, with a user guide and misuse reporting channels provided. **Social Benefits and Limitations:** This work curbs harmful content spread, protects vulnerable groups and promotes inclusive digital spaces. However, we acknowledge that no detection model can achieve 100% accuracy: false positives may restrict legitimate speech, and false negatives may allow harmful content to evade detection. Future work will focus on improving cross-cultural adaptability and reducing such risks.

References

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, and 1 others. 2025. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Rui Cao, Ming Shan Hee, Adriel Kuek, Wen-Haw Chong, Roy Ka-Wei Lee, and Jing Jiang. 2023. Procap: Leveraging a frozen vision-language model for hateful meme detection. In *Proceedings of the 31st ACM international conference on multimedia*, pages 5244–5252.

Rui Cao, Roy Ka-Wei Lee, Wen-Haw Chong, and Jing Jiang. 2022. Prompting for multimodal hateful meme classification. In *Proceedings of the 2022 conference on empirical methods in natural language processing*, pages 321–332.

Rui Cao, Roy Ka-Wei Lee, and Jing Jiang. 2024. Modularized networks for few-shot hateful meme detection. In *Proceedings of the ACM Web Conference 2024*, pages 4575–4584.

Ines Chami, Zhitao Ying, Christopher Ré, and Jure Leskovec. 2019. Hyperbolic graph convolutional neural networks. *Advances in neural information processing systems*, 32.

Anusha Chhabra and Dinesh Kumar Vishwakarma. 2023. Multimodal hate speech detection via multi-scale visual kernels and knowledge distillation architecture. *Engineering Applications of Artificial Intelligence*, 126:106991.

Karan Desai, Maximilian Nickel, Tanmay Rajpurohit, Justin Johnson, and Shanmukha Ramakrishna Vedantam. 2023. Hyperbolic image-text representations. In *International Conference on Machine Learning*, pages 7694–7731. PMLR.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186.

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

William Fedus, Barret Zoph, and Noam Shazeer. 2022. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research*, 23(120):1–39.

Elisabetta Fersini, Francesca Gasparini, Giulia Rizzi, Aurora Saibene, Berta Chulvi, Paolo Rosso, Alyssa Lees, and Jeffrey Sorensen. 2022. Semeval-2022 task 5: Multimedia automatic misogyny identification. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 533–549.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Liam Hebert, Gaurav Sahu, Yuxuan Guo, Nanda Kishore Sreenivas, Lukasz Golab, and Robin Cohen. 2024. Multi-modal discussion transformer: Integrating text, images and graph transformers to detect hate speech on social media. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 22096–22104.

Ming Shan Hee and Roy Ka-Wei Lee. 2025. Demystifying hateful content: Leveraging large multimodal models for hateful meme detection with explainable decisions. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 19, pages 774–785.

Paulo Cezar de Q Hermida and Eulanda M dos Santos. 2023. Detecting hate speech in memes: a review. *Artificial Intelligence Review*, 56(11):12833–12851.

Jinfa Huang, Jinsheng Pan, Zhongwei Wan, Hanjia Lyu, and Jiebo Luo. 2025. Evolver: Chain-of-evolution prompting to boost large multimodal models for hateful meme detection. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 7321–7330.

Junhui Ji, Xuanrui Lin, and Usman Naseem. 2024. Capalign: Improving cross modal alignment via informative captioning for harmful meme detection. In *Proceedings of the ACM Web Conference 2024*, pages 4585–4594.

Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. 2020. The hateful memes challenge: Detecting hate speech in multimodal memes. *Advances in neural information processing systems*, 33:2611–2624.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213.

Jian Lang, Rongpei Hong, Ting Zhong, Leiting Chen, Qiang Gao, and Fan Zhou. 2025. From shallow humor to metaphor: Towards label-free harmful meme detection via lmm agent self-improvement. *arXiv preprint arXiv:2512.21598*.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR.

Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. 2021. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, 34:9694–9705.

Liunian Harold Li, Mark Yatskar, D Yin, CJ Hsieh, and KW Chang. 2019. Visualbert: A simple and performant baseline for vision and language. arxiv. *arXiv preprint arXiv:1908.03557*, 10.

Hongzhan Lin, Ziyang Luo, Wei Gao, Jing Ma, Bo Wang, and Ruichao Yang. 2024. Towards explainable harmful meme detection through multimodal debate between large language models. In *Proceedings of the ACM Web Conference 2024*, pages 2359–2370.

Xuanrui Lin, Chao Jia, Junhui Ji, Hui Han, and Usman Naseem. 2025. Ask, acquire, understand: A multimodal agent-based framework for social abuse detection in memes. In *Proceedings of the ACM on Web Conference 2025*, pages 4734–4744.

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

Junxi Liu, Yanyan Feng, Jiehai Chen, Yun Xue, and Fenghuan Li. 2024a. Prompt-enhanced network for hateful meme classification. *arXiv preprint arXiv:2411.07527*.

Kangzheng Liu, Feng Zhao, Yu Yang, and Guandong Xu. 2024b. Dysarl: Dynamic structure-aware representation learning for multimodal knowledge graph reasoning. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 8247–8256.

Yifan Liu, Yaokun Liu, Zelin Li, Ruichen Yao, Yang Zhang, and Dong Wang. 2025. Modality interactive mixture-of-experts for fake news detection. In *Proceedings of the ACM on Web Conference 2025*, pages 5139–5150.

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

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 32.

Junyu Lu, Bo Xu, Xiaokun Zhang, Hongbo Wang, Haohao Zhu, Dongyu Zhang, Liang Yang, and Hongfei Lin. 2024. Towards comprehensive detection of chinese harmful memes. *Advances in Neural Information Processing Systems*, 37:13302–13320.

Paolo Mandica, Luca Franco, Konstantinos Kallidromitis, Suzanne Petryk, and Fabio Galasso. 2024. Hyperbolic learning with multimodal large language models. In *European Conference on Computer Vision*, pages 382–398. Springer.

Jingbiao Mei, Jinghong Chen, Weizhe Lin, Bill Byrne, and Marcus Tomalin. 2024. Improving hateful meme detection through retrieval-guided contrastive learning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5333–5347.

Gabriel Moreira, Manuel Marques, João Paulo Costeira, and Alexander Hauptmann. 2024. Hyperbolic vs euclidean embeddings in few-shot learning: Two sides of the same coin. In *Proceedings of the IEEE/CVF Winter conference on applications of computer vision*, pages 2082–2090.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

Avik Pal, Max van Spengler, Guido Maria D'Amely di Melendugno, Alessandro Flaborea, Fabio Galasso, and Pascal Mettes. 2024. Compositional entailment learning for hyperbolic vision-language models. *arXiv preprint arXiv:2410.06912*.

Someen Park, Jaehoon Kim, Seungwan Jin, Sohyun Park, and Kyungsik Han. 2024. Predict: multi-agent-based debate simulation for generalized hate speech detection. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 20963–20987.

Zelin Peng, Zhengqin Xu, Zhilin Zeng, Changsong Wen, Yu Huang, Menglin Yang, Feilong Tang, and Wei Shen. 2025. Understanding fine-tuning clip for open-vocabulary semantic segmentation in hyperbolic space. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 4562–4572.

Shraman Pramanick, Dimitar Dimitrov, Rituparna Mukherjee, Shivam Sharma, Md Shad Akhtar, Preslav Nakov, and Tanmoy Chakraborty. 2021. Detecting harmful memes and their targets. *arXiv preprint arXiv:2110.00413*.

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

Amanpreet Singh, Ronghang Hu, Vedanuj Goswami, Guillaume Couairon, Wojciech Galuba, Marcus Rohrbach, and Douwe Kiela. 2022. Flava: A foundational language and vision alignment model. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15638–15650.

Ilya Stepanov, Junaid Rashid, Jong Weon Lee, Salman Naseem, and Imran Razzak. 2025. Cbrcl: A clip-bert with retrieval-guided contrastive learning multimodal approach for crisis-driven hate speech detection. In *Companion Proceedings of the ACM on Web Conference 2025*, pages 1993–1999.

Abraham A Ungar. 2001. Hyperbolic trigonometry and its application in the poincaré ball model of hyperbolic geometry. *Computers & Mathematics with Applications*, 41(1-2):135–147.

Hans Van Ditmarsch, Wiebe van Der Hoek, and Barteld Kooi. 2008. *Dynamic epistemic logic*. Springer.

Ivan Vendrov, Ryan Kiros, Sanja Fidler, and Raquel Urtasun. 2015. Order-embeddings of images and language. *arXiv preprint arXiv:1511.06361*.

Bin Wang, Fuyong Xu, Peiyu Liu, and Zhenfang Zhu. 2024. Hypermr: Hyperbolic hypergraph multi-hop reasoning for knowledge-based visual question answering. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 8505–8515.

Bo Xu, Erchen Yu, Jiahui Zhou, Hongfei Lin, and Linlin Zong. 2025. Hyperhateprompt: A hypergraph-based prompting fusion model for multimodal hate detection. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 3825–3835.

Chuanpeng Yang, Fuqing Zhu, Yaxin Liu, Jizhong Han, and Songlin Hu. 2024. Uncertainty-aware cross-modal alignment for hate speech detection. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 16973–16983.

Jiayi Ye, Yanbo Wang, Yue Huang, Dongping Chen, Qihui Zhang, Nuno Moniz, Tian Gao, Werner Geyer, Chao Huang, Pin-Yu Chen, and 1 others. 2024. Justice or prejudice? quantifying biases in llm-as-a-judge. *arXiv preprint arXiv:2410.02736*.

Xianbing Zhao, Yixin Chen, Wanting Li, Lei Gao, and Buzhou Tang. 2022. Mag+: An extended multimodal adaptation gate for multimodal sentiment analysis. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4753–4757. IEEE.

Changmeng Zheng, Dayong Liang, Wengyu Zhang, Xiao-Yong Wei, Tat-Seng Chua, and Qing Li. 2024. A picture is worth a graph: A blueprint debate paradigm for multimodal reasoning. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 419–428.

Hengyang Zhou, Jinwu Yan, Yaqing Chen, Rongman Hong, Wenbo Zuo, and Keyan Jin. 2025. Ldgnet: Llms debate-guided network for multimodal sarcasm detection. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

A Related Work

A.1 Detection of Harmful Memes

Harmful memes have proliferated extensively across social media platforms, inflicting harm on vulnerable populations and contributing to increased social fragmentation to some degree. To address the challenge of effective detection, (Stepanov et al., 2025) employs retrieval-guided contrastive learning to improve hate speech identification. (Yang et al., 2024) introduces a cross-modal alignment framework to model multimodal error alignment and uncertainty perception. (Chhabra and Vishwakarma, 2023) utilizes a multi-scale adaptive receptive field to emphasize salient spatial regions, while (Xu et al., 2025) leverages hypergraphs to capture the hateful content arising from cross-modal information. Additionally, (Lin et al., 2024) implements modular networks for the detection of hate memes. Although these methods primarily focused on directly training classifiers for hate meme detection, they often overlooked the supplementary role of implicit information embedded within the memes. In contrast, studies such as (Ji et al., 2024; Park et al., 2024) directly apply VLMs to extract indicative features from memes. Furthermore, (Lin et al., 2024; Huang et al., 2025; Hee and Lee, 2025; Liu et al., 2024a) propose that VLMs be employed to extract implicit information from memes, which is then integrated into detection models during training. Distinct from the aforementioned approaches, our method introduces a novel concept termed "Viewpoint Decoupling". This approach has the potential to mitigate the inherent biases of models toward harmful content to a certain extent.

A.2 Hyperbolic Space

Hyperbolic space can effectively model data with potential hierarchical structures, thereby enhancing the generalization ability of models. (Liu et al., 2024b) utilizes hyperbolic space to solve the task of completing multimodal knowledge graphs and designs a dual-space multi-hop structure learning module. (Mandica et al., 2024) proposes a BLIP-2 hyperbolic version training strategy. (Moreira et al., 2024) demonstrates that hyperbolic embeddings achieve the best few-shot classification performance. (Peng et al., 2025) proposes HyperCLIP, which fine-tunes text embeddings by adjusting their hyperbolic radius through scaling transformations. (Desai et al., 2023; Pal et al., 2024) prove the existence of a hierarchical structure between image-text pairs. We posit the existence of a hierarchical relationship between multimodal information and interpretative perspectives, and draw upon the aforementioned concepts to model this hierarchical structure within hyperbolic space.

Harmful Prompt

Given an image and its corresponding title, please explain, from the perspective of hatred or harm, why this image and caption have been labeled as hateful or harmful.
Please note that you need to provide an explanation from a hateful and harmful perspective, and keep it within no more than two sentences.
Here is the title:{title}
Your answer:

Non-harmful Prompt

Given an image and its corresponding title, please explain, from a non-hateful or harmless perspective, why this image and caption have been labeled as non-hateful or harmless.
Please note that you need to provide the explanation from a non-hateful and harmless perspective, and keep it within no more than two sentences.
Here is the title:{title}
Your answer:

Figure 8: A prompt used to generate different perspectives.

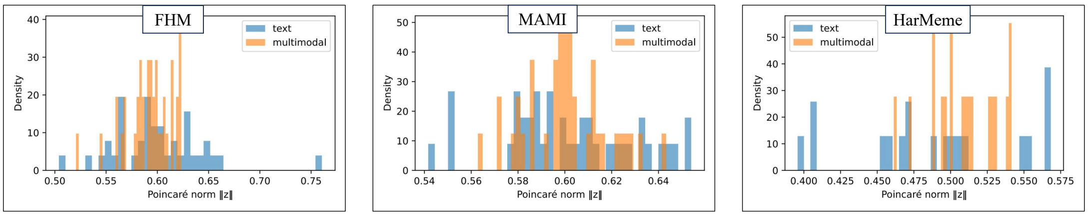
B Viewpoint Generation Template

To facilitate the model's ability to produce viewpoints from two distinct perspectives, we meticulously crafted a prompt that includes instructions representing both a harmful and a harmless viewpoint, as illustrated in Figure 8. It is important to note that the generated viewpoints were restricted to a maximum of two sentences, taking into account the computational constraints of the visual language model.

C Loss Calculation

To enhance the discrimination between positive and negative samples within our framework, we incorporate a contrastive learning loss. Let the total number of samples be denoted by n, and suppose that among these, there are m positive samples sharing the same label. The feature representations of all samples are given by the set $\{f_1, f_2, \ldots, f_n\}$. Under these conditions, the following formulation applies:

$$\mathcal{L}_{InfoICE} = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{m}\sum_{j=1}^{m}\log\frac{\exp(s_{i,j})}{\sum\limits_{k:k\neq i}^{n}\exp(s_{i,k})} \tag{16}$$



(a) (b) (c)

Method	AUC ($\downarrow$0.84)	ACC ($\downarrow$0.6)
BPDMoE-Hate	83.71±0.39	75.18±1.38
No Hierarchy	82.87±0.56	74.58±1.46

Method	AUC ($\uparrow$0.46)	ACC ($\downarrow$1.12)
BPDMoE-Hate	87.84±0.54	76.70±0.80
No Hierarchy	88.30±0.47	75.58±1.49

Method	AUC ($\downarrow$0.73)	ACC ($\downarrow$1.02)
BPDMoE-Hate	94.11±0.28	86.10±0.94
No Hierarchy	93.38±0.37	85.08±1.57

Figure 9: Eliminate the hierarchical relationship between viewpoints and multimodal features in hyperbolic space.

Here, $s_{i,j} = f_i^T f_j / \tau$, where $\tau = 0.07$ is the temperature coefficient. Moreover, to ensure the balanced loading of experts, our load balancing loss is as follows:

$$\mathcal{L}_{balance} = \frac{1}{n}\sum_{i=1}^{n}(w_{\text{exp}}^{i} - \bar{w}_{\text{exp}})^2 + \frac{1}{n}\sum_{i=1}^{n}(c_{\text{exp}}^{i} - \bar{c}_{\text{exp}})^2 \quad (17)$$

Where n represents the number of experts, w_{exp}^i represents the weight assigned to the ith expert, and c_{exp}^i represents the loading frequency of the ith expert in a sampling set.

Dataset	FHM	MAMI	HarMeme
Train	8,500	10,000	3,013
Valid	500	100	177
Test	1,000	1,000	354

Table 5: Dataset Distribution

D Detailed Supplement of the Dataset and Baselines

We conducted an evaluation of BPDMoE-Hate utilizing three extensively recognized hate meme datasets. The FHM dataset comprises a diverse collection of harmful memes sourced from the internet, consisting of 8,500 samples designated for training and 1,000 samples reserved for testing. The MAMI dataset focuses specifically on misogynistic memes gathered from prominent social media platforms. The HarMeme dataset includes harmful memes obtained from social media websites as well as through crowdsourcing efforts, notably encompassing a substantial subset of memes associated with the 2019 novel coronavirus. The size of the relevant dataset is shown in table 5.

The models we used for comparison include: 1) VLMs, such as Qwen2.5-vl-Instruct-32B (Bai et al., 2025), Llama-3.2-11B-Vision (Grattafiori et al., 2024) and Llava-1.5 (Liu et al., 2023). 2) Pure text classification models, including Bert-base (Devlin et al., 2019) and RoBERTa-large (Liu et al., 2019). 3) Multi-modal classification models, including FLAVA-full (Singh et al., 2022), Visual-BERT (Li et al., 2019), ViLBERT (Lu et al., 2019), BLIP2 (Li et al., 2023) and ALBEF (Li et al., 2021). 4) Harmful meme detection frameworks, including Mod-HATE (Cao et al., 2024), PromptHate (Cao et al., 2022), Pro-Cap (Cao et al., 2023), Explain-HM (Lin et al., 2024) and IntMeme (Hee and Lee, 2025).

E Additional Experiments

E.1 The Influence of the Viewpoint Encoder

For the encoding of viewpoints, we employed the more advanced RoBERTa-large model as the text encoder. This section examines the influence of various viewpoint encoders on the overall model performance, as shown in Table 6. Specifically, we substituted the original encoder with four alternative models: RoBERTa-base, T5-base (Raffel et al., 2020), BERT-base, and BERT-large. Experimental results indicate that an increase in the number of model parameters does not necessarily correspond to improved performance in meme detection. Rather, the effectiveness appears to depend on the intrinsic capabilities of the encoder itself. We hypothesize that this outcome may be attributed to the extensive freezing of parameters during the training phase. Consequently, future work should focus on selecting a more effective text encoder to further enhance the framework's performance.

View Encoder	FHM		MAMI		HarMeme	
	AUC	ACC	AUC	ACC	AUC	ACC
Roberta-large	83.71 $\pm$0.39	75.18 $\pm$1.38	87.84 $\pm$0.54	76.70 $\pm$0.80	94.11 $\pm$0.28	86.10 $\pm$0.94
Roberta-base	83.89 $\pm$0.49	75.06 $\pm$1.20	88.91 $\pm$0.48	75.40 $\pm$1.06	93.85 $\pm$0.43	86.05 $\pm$1.48
T5-base	83.11 $\pm$0.74	73.92 $\pm$0.72	88.21 $\pm$0.36	75.02 $\pm$1.53	93.23 $\pm$0.32	84.40 $\pm$1.75
Bert-large	83.14 $\pm$0.92	73.28 $\pm$2.12	88.63 $\pm$0.32	76.90 $\pm$2.21	92.73 $\pm$0.26	84.92 $\pm$1.14
Bert-base	83.49 $\pm$0.58	73.72 $\pm$0.76	89.05 $\pm$0.52	76.46 $\pm$0.97	93.78 $\pm$0.30	85.54 $\pm$0.70

Table 6: The impact of the viewpoint encoder (text encoder) on the model's performance.

Multimodal Encoder	FHM		MAMI		HarMeme	
	AUC	ACC	AUC	ACC	AUC	ACC
BLIP2	83.71 $\pm$0.39	75.18 $\pm$1.38	87.84 $\pm$0.54	76.70 $\pm$0.80	94.11 $\pm$0.28	86.10 $\pm$0.94
ViT	80.83 $\pm$0.37	71.94 $\pm$0.86	85.98 $\pm$0.42	74.20 $\pm$0.63	85.95 $\pm$1.49	79.15 $\pm$1.12
Flava-full	83.34 $\pm$0.67	74.50 $\pm$0.87	87.40 $\pm$0.70	75.74 $\pm$1.51	89.62 $\pm$0.77	83.33 $\pm$0.67

Table 7: The impact of the multimodal encoder on the model's performance.

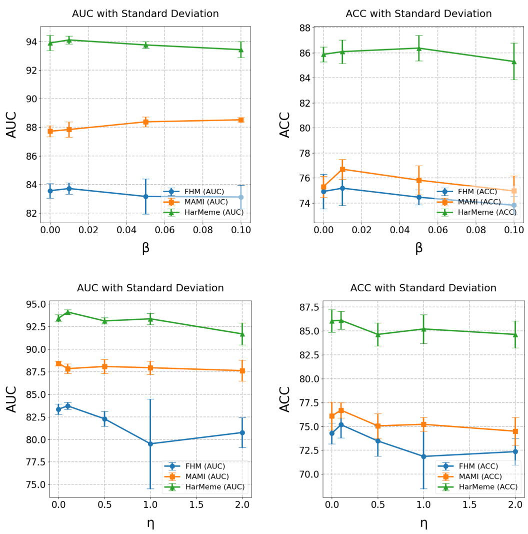


Figure 10: Compare the influence of the coefficients of the contrastive loss and the entailment loss. "β" represents the coefficient of the entailment loss, and "η" indicates the coefficient of the contrastive loss.

E.2 The Influence of the Multimodal Encoder

The performance of the multimodal encoder directly influences the feature representations during the hierarchical fusion stage. Since both images and their corresponding captions are simultaneously input into the multimodal encoder, the features output by this encoder have a greater impact on the BPDMoE-Hate prediction. As shown in Table 7, compared to Table 1, employing the lower-performing Flava-full model results in a significant decline in overall model performance. This finding underscores the critical role of multimodal information as the "root" within the hierarchical structure; inadequate representation of this information severely impairs hierarchical modeling. Furthermore, we replaced the multimodal encoder with a ViT (Dosovitskiy, 2020) model capable of encoding only images to investigate whether a more effective hierarchical structure could be formed between "image" and "viewpoint". Experimental results indicate that this hierarchical structure contributes less to our model than the "multimodal-viewpoint" structure, thereby further validating the feasibility of the proposed hierarchical framework.

E.3 Different Viewpoint Generation Models

Our BPDMoE-Hate framework employs the advanced Qwen2.5-VL-32B-Instruct model for viewpoint generation. In this section, we investigate the influence of utilizing various VLMs to produce binary viewpoints on the overall system performance. The evaluation was conducted on the FHM and HarMeme datasets, with the corresponding results presented in table 8. Our findings indicate that VLMs possessing stronger self-inference capabilities exert a more beneficial impact on the framework. Conversely, VLMs with comparatively weaker performance, constrained by limited internal knowledge, tend to generate binary viewpoints with reduced informational content, thereby impeding the model's evaluative accuracy. Consequently, we infer that binary viewpoints derived from models with enhanced reasoning abilities more accurately capture the authentic expression of memes, leading to improved judgment within the framework.

VLM	FHM		HarMeme	
	AUC	ACC	AUC	ACC
Qwen2.5-VL-32B-Instruct	83.71 $\pm$0.39	75.18 $\pm$1.38	94.11 $\pm$0.28	86.10 $\pm$0.94
Gemma3-12B	84.37 $\pm$0.55	75.50 $\pm$1.28	93.88 $\pm$0.59	83.95 $\pm$1.84
Qwen3-VL-8B	82.65 $\pm$0.59	72.44 $\pm$1.14	94.16 $\pm$0.29	84.41 $\pm$1.20
Qwen2.5-VL-7B-Instruct	80.48 $\pm$0.41	69.66 $\pm$2.65	92.74 $\pm$0.51	84.01 $\pm$1.91
Qwen2-VL-2B-Instruct	74.47 $\pm$0.34	64.82 $\pm$1.98	92.59 $\pm$1.22	83.90 $\pm$1.83

Table 8: The influence of different viewpoint generation models on the results.

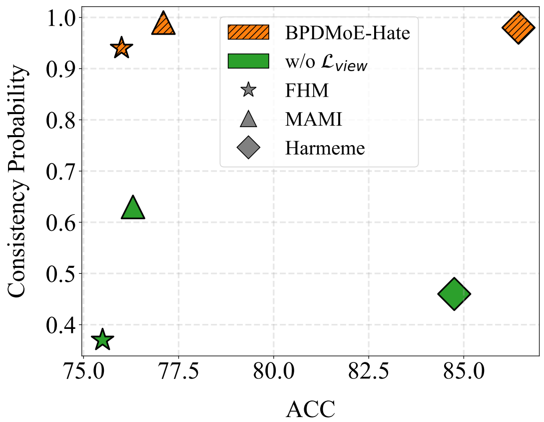


Figure 11: Removing the viewpoint selection loss. The term "Consistency Probability" denotes the probability that the model's predicted viewpoint selection aligns with the harmful meme prediction within the test dataset. We fix the random seed for the verification.

E.4 The Proportion of Different Losses

This section examines the effects of the entailment loss and the contrastive learning loss on model performance. Since the viewpoint selection loss and the task loss are components of the same optimization objective, and no discernible pattern was observed regarding the impact of the corresponding loss coefficient on the outcomes during experimentation, further discussion on this aspect is omitted. The experimental results are presented in Figure 10. Our findings indicate that optimal performance was achieved at parameter values of $\beta = 0.01$ and $\eta = 0.1$. Although increasing β led to a slight improvement in the AUC metric for FHM, the ACC of HarMeme peaked at $\eta = 0.05$. Furthermore, we hypothesize that the proportion of entailment loss should not be excessively large because it is applied exclusively within the hyperbolic space. An overly high weighting of this loss may inhibit the model's ability to select experts effectively in the hyperbolic space, thereby diminishing overall performance.

E.5 Consistency Between Viewpoint Selection and Prediction

In this section, we remove the viewpoint selection loss to examine the consistency between the model's viewpoint selection predictions and its harmful meme predictions. The results are presented in the figure 11. We observe that after training, BPDMoE-Hate demonstrates a high degree of alignment between viewpoint selection and the prediction of whether a meme is harmful, underscoring the significance of the viewpoint selection loss for our model. Upon removal of this loss, the "Consistency Probability" experiences a substantial decline across all three datasets, and the ACC of the model is inferior to that of BPDMoE-Hate. This indicates that accurate viewpoint selection positively influences model performance, whereas incorrect viewpoint choices tend to mislead harmful meme predictions. These findings further validate the efficacy of our model in mitigating bias.

Type	Parameter Size
Trainable params	138 M
Non-trainable params	4.1 B
Total Params	4.2 B

Table 9: The total number of parameters.

E.6 Eliminate Hierarchical Relationships

In this section, we verify the impact of eliminating the hierarchical relationship in hyperbolic space on the model's performance. We artificially bring the norm values of the projected viewpoint features and multimodal features in hyperbolic space closer and observe the experimental results as shown in Figure 9. We find that the absence of hierarchical relationship constraints leads to a certain decline in model performance. Moreover, the performance metrics in Figure 6 decline more significantly compared to the elimination of hierarchical relationships, demonstrating the importance of correctly setting the hierarchical relationship order.

Hyperbolic Distance	FHM	MAMI	HarMeme
Multimodal	0.63	0.60	0.67
Viewpoint	7.82	8.57	7.17
Difference Value	7.19	7.97	6.50

Table 10: The average hyperbolic distance of different features from the origin. Here, we fixed the random seed for the test set.

E.7 Quantitative Analysis of Hierarchical Structure

The preceding observations have been intuitively illustrated through the distribution histograms of the norms of various vectors, indicating the existence of a hierarchical relationship between multimodal features and viewpoint features. Furthermore, multimodal features represent a more generalized form compared to viewpoint features. In this section, we quantitatively validate this hierarchical relationship by analyzing the average hyperbolic distance from the origin for both multimodal and viewpoint features within different test sets. The results, presented in table 10, demonstrate that across different test sets, multimodal features consistently exhibit shorter hyperbolic distances to the origin, whereas the selected viewpoint features are positioned farther away. These findings substantiate that the proposed model effectively captures the hierarchical structure inherent between these two feature types. It is observed that the range of values for the hyperbolic distance is $[0, +\infty)$.

Dataset	BPDMoE-Hate	w/o VEM
FHM	75.18 $\pm$1.38	71.74 $\pm$0.72
MAMI	76.70 $\pm$0.80	74.04 $\pm$0.55
HarMeme	86.10 $\pm$0.94	83.05 $\pm$0.89

Table 11: The significance of the viewpoint enhancement module. The evaluation metric used is ACC.

E.8 Removal of The Viewpoint Enhancement Module

The design objective of the VEM is to comprehensively incorporate the essential characteristics of contrasting perspectives and multimodal data into the target viewpoint, thereby better capturing the harmful information hidden in memes. Consequently, this module constitutes a critical component of our overall framework. To evaluate its

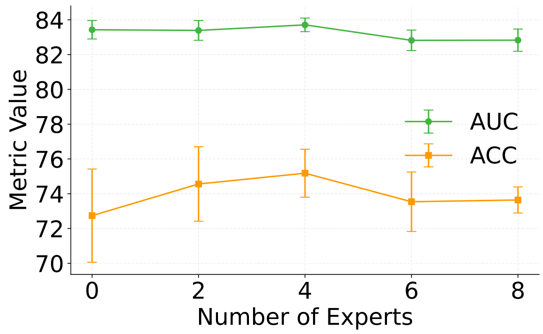


Figure 12: The impact of the number of experts. This assessment is conducted using the FHM dataset.

significance, we conducted an ablation study by removing the VEM and assessing the resultant performance variations. The findings, as presented in table 11, indicate a performance decline following the module's removal. Notably, the VEM exerted the most pronounced effect on the FHM dataset and the least on the MAMI dataset. This observation further substantiates our assertion that the MAMI dataset, which focuses on misogyny, possesses relatively straightforward discriminative features, wherein the integration of images and titles sufficiently conveys explicit information.

E.9 Number of Experts

The number of selected experts is fixed at 2, while the total number of experts is varied. The corresponding results are presented in the Figure 12. It is important to note that a total expert count of zero indicates the direct removal of experts from the dual space. The findings demonstrate that the performance of BPDMoE-Hate is influenced by the total number of experts. Specifically, optimal performance is observed when the number of experts was set to 4. Consequently, the total number of experts is established at 4 for subsequent experiments.

E.10 Text-only Probe on Generated Viewpoint Features

In this section, we employ the RoBERTa-large model trained exclusively on the generated textual content, without incorporating any visual information. As presented in Table 12, even when leveraging both viewpoints, the text-only classifier still lags considerably behind our full model (HarMeme AUC: 88.39 vs. 94.11), demonstrating that visual features and hierarchical fusion are essential for strong performance. Notably, adopting

Method	FHM		MAMI		HarMeme	
	AUC	ACC	AUC	ACC	AUC	ACC
Text-only (Non-harmful)	78.42 $\pm$0.29	68.72 $\pm$1.41	78.45 $\pm$0.54	70.76 $\pm$1.09	83.68 $\pm$0.44	77.23 $\pm$0.96
Text-only (Harmful)	81.84 $\pm$0.21	72.30 $\pm$0.26	84.06 $\pm$0.16	73.38 $\pm$0.78	86.74 $\pm$0.56	79.44 $\pm$0.95
Text-only (Both)	83.50 $\pm$0.47	72.88 $\pm$0.52	84.81 $\pm$0.16	74.20 $\pm$0.89	88.39 $\pm$0.53	82.20 $\pm$0.64
BPDMoE-Hate (Full)	**83.71** $\pm$**0.39**	**75.18** $\pm$**1.38**	**87.84** $\pm$**0.54**	**76.70** $\pm$**0.80**	**94.11** $\pm$**0.28**	**86.10** $\pm$**0.94**

Table 12: Text-only Probe Experiments on Generated Viewpoint Features

Method	FHM	MAMI	HarMeme
Image + Harmful viewpoint (only vlm)	71.32 $\pm$0.26	75.70 $\pm$1.79	55.48 $\pm$0.23
Image + Non-harmful viewpoint (only vlm)	60.34 $\pm$0.08	53.52 $\pm$0.04	53.52 $\pm$0.04
Image + Both viewpoints (only vlm)	71.22 $\pm$1.72	70.20 $\pm$1.67	67.23 $\pm$0.00
BPDMoE-Hate	**75.18** $\pm$**1.38**	**76.70** $\pm$**0.80**	**86.10** $\pm$**0.94**

Table 13: Direct VLM Decision Baselines for Image-Viewpoint Inputs. The evaluation metric used is ACC.

dual viewpoints instead of relying on a single best viewpoint yields AUC improvements of 1.66% on FHM, 0.75% on MAMI, and 1.65% on HarMeme. This confirms that our adversarial generation strategy remains beneficial even in the absence of visual modalities.Nevertheless, the performance upper bound of the text-only model (around 88 AUC on HarMeme) is clearly surpassed by our DSMoE-based hierarchical fusion approach (94.11 AUC). This validates that geometric representation learning effectively captures meaningful image–text interactions that cannot be derived from raw textual semantics alone.

E.11 Direct Classification Baselines with VLM

We further conduct an end-to-end classification experiment by feeding images and their generated captions directly into Qwen2.5-VL. As reported in Table 13, the results highlight clear limitations of vanilla vision-language model outputs. When using only harmful viewpoints, classification accuracy reaches 71.32% on FHM and 75.70% on MAMI, yet drops sharply to 55.48% on HarMeme. For non-harmful viewpoints, performance hovers near random guessing across all datasets, ranging from 53% to 60%. Simply concatenating both viewpoints without explicit selection or fusion yields accuracies of 71.22%, 70.20%, and 67.23%—still significantly lower than the 75.18%, 76.70%, and 86.10% achieved by BPDMoE-Hate. This indicates that unfiltered dual perspectives can interfere destructively with prediction. In contrast, our proposed AVG+DSMoE modules deliver substantial practical benefits: absolute accuracy improvements of 3.96%, 6.50%, and 18.87% over the strongest vanilla VLM baseline.

E.12 Human Evaluation of Model Interpretability

Metric	Value
Human assessment accuracy rate	0.820
AVG quality evaluation (average)	3.99
AVG accuracy rate (all data)	0.748
AVG accuracy rate (sample data)	0.640

Table 14: Systematic human evaluation results for model interpretability (FHM dataset, 50 sampled examples).

To further validate the interpretability of BPDMoE-Hate in a rigorous and quantitative manner, we conduct a systematic human evaluation on the FHM dataset (fixed random seed 42), including assessments of viewpoint quality, Adaptive Viewpoint Gating (AVG) selection accuracy, and failure mode analysis for AVG errors. For the experiment, we uniformly sample 50 representative examples from the FHM dataset, invite a professional evaluator in this field, and design a three-stage human evaluation protocol: (1) human judgment accuracy (annotators are provided with dual viewpoints and blind to true labels); (2) human quality rating of AVG-selected viewpoints (on a 1–5 scale, with 5 indicating the highest quality); (3) AVG selection accuracy (consistency between AVG's selected viewpoint and the true label of the meme).

The quantitative results of the human evaluation and AVG performance are summarized in Table 14. We find that the average quality score of AVG-selected viewpoints reaches 3.99, approach-

Type of error	Num	Typical characteristic
Visual neglect	1	Image contains implicit hate symbols
Ambiguity	6	Both viewpoints are reasonable
Deficiency of context	7	Requires cultural background knowledge
Poor quality	4	Low-quality viewpoints generated by VLM

Table 15: Failure mode analysis of AVG selection errors.

ing human-level judgment, while the AVG selection accuracy on sampled data (64.0%) lags significantly behind the human baseline (82.0%). This indicates that the quality of VLM-generated viewpoints is the core bottleneck for interpretability performance, rather than the AVG selection mechanism itself, which further validates a key limitation of our framework: its partial dependence on the quality of dual viewpoints generated by VLMs.

We further conduct a fine-grained failure mode analysis on the 18 misjudgments (36% of the 50 sampled examples) of the AVG module, and categorize the error types and their typical characteristics in Table 15.

The analysis shows that AVG selection errors mainly stem from two core factors: inadequate VLM generation quality and lack of cultural background knowledge in viewpoint generation. Corresponding future improvement solutions include adopting more powerful VLMs for viewpoint generation and integrating retrieval-augmented generation (RAG) to supplement contextual and cultural knowledge.

F The Number of Parameters

The total number of parameters within the proposed framework was computed and is presented in table 9. It is important to note that the parameters associated with the VLM were excluded from this calculation. The framework comprises 138 million trainable parameters, whereas the majority of non-trainable parameters originate from the multimodal encoder component. Substituting the multimodal encoder with a more lightweight alternative, such as FLAVA-full (as shown in table 7), enables deployment on a wider range of devices; however, this modification will incur a slight reduction in performance.

G Deployment Efficiency Analysis

We measure the inference time of BPDMoE-Hate using 1,000 samples as an example, with the results shown in the Table 16. The results indicate that our method performs inference quickly on individual samples, enabling efficient deployment.

Time statistics for reasoning	
Total number of test samples	1,000
Total reasoning time	91.6237s
Average processing time per data	91.62ms

Table 16: Time complexity statistics of our method.