PRISM of Opinions: A Persona-Reasoned Multimodal Framework for User-centric Conversational Stance Detection

Anonymous ACL submission

Abstract

The rapid proliferation of multimodal social media content has driven research in Multimodal Conversational Stance Detection (MCSD), which aims to interpret users' attitudes toward specific targets within complex discussions. However, existing studies remain limited by: **1) pseudo-multimodality**, where visual cues appear only in source posts while comments are treated as text-only, misaligning with real-world multimodal interactions; and **2) user homogeneity**, where diverse users are treated uniformly, neglecting personal traits that shape stance expression. To address these issues, we introduce **U-MStance**, the first user-centric MCSD dataset, containing over 40k annotated comments across six real-world targets. We further propose **PRISM**, a **P**ersona-**R**easoned mult**I**modal **S**tance **M**odel for MCSD. PRISM first derives longitudinal user personas from historical posts and comments to capture individual traits, then aligns textual and visual cues within conversational context via Chain-of-Thought to bridge semantic and pragmatic gaps across modalities. Finally, a mutual task reinforcement mechanism is employed to jointly optimize stance detection and stance-aware response generation for bidirectional knowledge transfer. Experiments on U-MStance demonstrate that PRISM yields significant gains over strong baselines, underscoring the effectiveness of user-centric and context-grounded multimodal reasoning for realistic stance understanding.

1 Introduction

Social media have become a dominant platform for the expression of opinions, making their analysis crucial to understand divisive societal topics such as politics, public health, and social justice (Glandt et al., 2021; Liang et al., 2022). Stance detection, the task of identifying a user's position toward a target, has become vital for charting public discourse and polarization. As online interactions grow more complex, the field has evolved from simple textual analysis to modeling rich, multi-party conversational contexts (Li et al., 2023b,d; Niu et al., 2024).

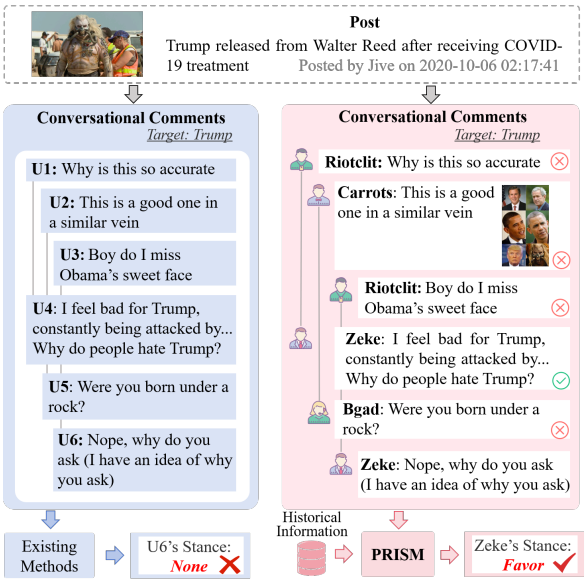

Figure 1: Comparison of conventional MCSD (left) with our proposed PRISM framework (right).

The recent shift towards Multimodal Conversational Stance Detection (MCSD), exemplified by the MmMtCSD dataset (et. al, 2024), marks a crucial step forward. However, despite this progress, two fundamental limitations persist across the field, hindering authentic stance modeling. The first is a **pseudo-multimodality**, where visual cues appear only in source posts while comments remain text-only, diverging from the multimodal nature of real-world interactions. The second is **user homogeneity**, where diverse individuals are modeled uniformly, ignoring personal traits that shape stance expression. These limitations can be clearly observed in the example shown in Figure 1. User *Carrots* combines an image and textual content to contrast Obama and Trump, conveying a derogatory attitude toward Trump, which highlights the necessity of multimodal cues for accurate stance understanding. Meanwhile, *Zeke* expresses sympathy

for Trump, and *Bgad* challenges this view, revealing conflicting stances. A model that remains blind to their individual personas would fail to grasp the reason for their disagreement, leading to incorrect stance interpretations.

To address these limitations, we therefore introduce **U-MStance**, the first dataset designed for user-centric multimodal conversational stance detection. Comprising over 40k annotated comments across six real-world targets, U-MStance uniquely incorporates multimodality in both source posts and conversational replies, providing a realistic and challenging benchmark that captures the nuances of user-centric, multimodal interactions. Building upon this critical new dataset, we further propose **PRISM**, a **P**ersona-**R**easoned mult**I**modal **S**tance **M**odel for MCSD. Specifically, PRISM first explores the underlying personal tendencies behind stance expressions and derives longitudinal user persona traits inspired by Big Five (OCEAN) theory (Jiang et al., 2024) from historical posts and comments. Then, it introduces a rationalized cross-modal grounding module, leveraging Chain-of-Thought (CoT) reasoning to explicitly bridge the profound semantic and pragmatic gap between modalities in conversational contexts. Ultimately, these personas and contextual representations are fed into a mutual task reinforcement mechanism, which jointly optimizes stance classification and stance-aware response generation, enabling bidirectional knowledge transfer that enhances both detection performance and the modeling of user communication styles. Comprehensive experiments on our U-MStance dataset show that PRISM achieves significant improvements over strong baselines across multiple metrics. The main contributions can be summarized as follows:

- To our knowledge, **U-MStance** is the first user-centric multimodal conversational stance detection dataset, which extends multimodality beyond posts to comments, and incorporates user information for more realistic and challenging stance understanding.

- We propose **PRISM**, a framework that models user personas and deepens cross-modal correlations for improved user-centric stance detection.

- Extensive experiments conducted on our U-MStance dataset demonstrate that PRISM significantly outperforms strong baselines, confirming both its effectiveness and generalization ability.

2 Related Works

2.1 Stance Datasets

The development of stance detection datasets has evolved from analyzing isolated texts to modeling complex interactions. Early benchmarks like SemEval-2016 Task 6 (Mohammad et al., 2016) and VAST (Allaway and McKeown, 2020a) focused on sentence-level textual stance detection, establishing foundational models but overlooking the context of dialogue. This limitation spurred the creation of conversational datasets such as SRQ (Villa-Cox, 2020) and CANT-CSD (Li et al., 2023c), which treated stance as a dynamic process unfolding within a conversation. The most recent advancement is the shift toward multimodal stance detection, which incorporates visual elements alongside text to better reflect modern social media discourse. However, even the latest datasets for MCSD, such as MmMtCSD (et. al, 2024), still face two key shortcomings that hinder realistic stance modeling. First, multimodality is constrained to source posts as static background context, while conversational replies are restricted to text-only. In addition, information about users, including their posting history, is largely ignored, overlooking contextual factors that shape an individual's expressed stance and leaving a substantial gap in building comprehensive and authentic stance detection models.

2.2 Stance Detection Methods

Methodological advancements in stance detection have spanned multiple task settings in-target (Li and Caragea, 2021), cross-target (Ding et al., 2024a; Liang et al., 2021; Wei and Mao, 2019), and zero-shot scenarios (Allaway and McKeown, 2020b; Liang et al., 2022; Zhu et al., 2022). Early deep learning approaches primarily fine-tuned pre-trained models such as BERT (Cambria et al., 2018). However, the field was revolutionized by the emergence of Large Language Models (LLMs), whose powerful semantic understanding and adaptability introduced new capabilities (Cheng et al., 2024a,b; Li et al., 2024; Ding et al., 2024b).

Following this paradigm shift, LLM-centric research has moved beyond direct prompting or CoT reasoning toward increasingly sophisticated methodologies (Zhang et al., 2023, 2022). These include fine-tuning with external knowledge (Li

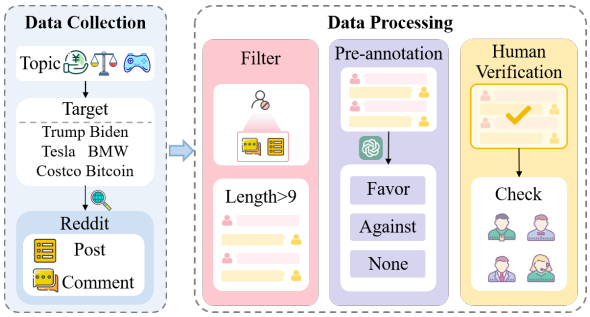

Figure 2: Data construction pipeline of U-MStance.

et al., 2023a), prompt-based multi-perspective reasoning (Ding et al., 2024a; Wang et al., 2025), formalizing logic structures (Zhang et al., 2024), and designing multi-agent collaborative systems (Lan et al., 2024; Wang et al., 2024). Despite this rapid progress, current approaches often treat users as homogeneous, overlooking the personal traits and discourse tendencies that shape stance expression. This limits the interpretation of personalized reactions and conflicting views, necessitating a shift from content-centered to user-centric modeling.

3 Dataset

This section outlines the construction and key attributes of our U-MStance dataset. As illustrated in Figure 2, our data construction pipeline includes collection, processing and verification to ensure data quality and relevance. Subsequently, data analysis is conducted to profile the dataset's composition and validate its quality.

3.1 Data Construction

Dataset Collection. To capture realistic user interactions, we sourced conversational data from Reddit[1], a large and diverse online forum, using the PRAW library. Covering the period from 2017 to 2025, we initially identified 835 posts related to six high-engagement, real-world targets: *Trump*, *Biden*, *Tesla*, *BMW*, *Costco*, and *Bitcoin*. These posts were selected based on popularity metrics like upvotes and comment counts to ensure a sufficient volume of conversational replies. We also collected each user's historical posts and comments to enrich their personal information.

Filter. To ensure quality and privacy, raw data was filtered by removing inactive accounts and threads exceeding depth 9, with all personally identifiable information replaced through a rigorous anonymization process.

[1]https://www.reddit.com

Table 1: Comparison of U-MStance with existing datasets. **Size**: the number of annotated instances; **Target**: the number of the targets; **Conv.**: support for the conversational context; **Vision-Related**: percentage of vision-integrated instances; **User**: support for user information; -: not applicable.

Dataset	Size	Target	Conv.	Vision-Related	User
SEM16 (Mohammad et al., 2016)	4,870	6	✗	✗	✗
P-stance (Li et al., 2021)	21,574	3	✗	✗	✗
COVID-19 (Glandt et al., 2021)	6,133	4	✗	✗	✗
WT-WT (Conforti et al., 2020)	51,284	5	✗	✗	✗
VAST (Allaway and McKeown, 2020a)	23,525	-	✗	✗	✗
SRQ (Villa-Cox, 2020)	1,348	4	✗	✗	✗
CANT-CSD (Li et al., 2023c)	5,376	1	✓	✗	✗
CTSDT (Li et al., 2023d)	53,861	1	✓	✗	✗
MT-CSD (Niu et al., 2024)	15,876	5	✓	✗	✗
C-MTCSD (Niu et al., 2025)	24,264	5	✓	✗	✗
MT^2-CSD (Niu, 2025)	24,457	6	✓	✗	✗
MMVax (Weinzierl and Harabagiu, 2023)	11,300	-	✗	100%	✗
MMSD (Liang et al., 2024)	17,544	5	✗	100%	✗
MmMtCSD (et. al, 2024)	21,340	3	✓	65.99%	✗
U-MStance(Ours)	40,003	6	✓	100%	✓

Pre-annotation. Following this, GPT-4o-mini was employed for automated pre-annotation, assigning a preliminary stance label (Favor, Against, or None) to each instance relative to its target.

Human Verification. Following qualification via two pilot rounds and an expert review, seven NLP researchers rigorously verified the automated labels using the interface illustrated in Appendix D. During the main annotation, each instance was independently reviewed by at least two annotators. Disagreements were resolved by a third senior annotator, and the final labels were determined through majority voting. Beyond stance correction, annotators also evaluated the contextual relevance of any associated images. This multi-annotator workflow ensures high label fidelity, and reinforces the overall reliability of the dataset.

3.2 Data Analysis

Data Statistics. Our dataset comprises 40,003 annotated instances and 24952 user profiles. Each instance consists of a complete conversational thread accompanied by its associated images within the conversation. Crucially, Table 1 provide a comparison between U-MStance and existing datasets. U-MStance is the first dataset in this domain to systematically incorporate user information, offering a foundational resource to address the prevalent issue of user homogeneity. More comprehensive statistics, including the distributions of labels and conversational depths, are detailed in Appendix A.

Quality Assessment. To ensure the annotation reliability, we measured inter-annotator agreement using Cohen's Kappa (Cohen, 1960), achieving an average score of 0.64, indicative of substantial agreement. This consistency was further reinforced through a rigorous post-annotation review, in which

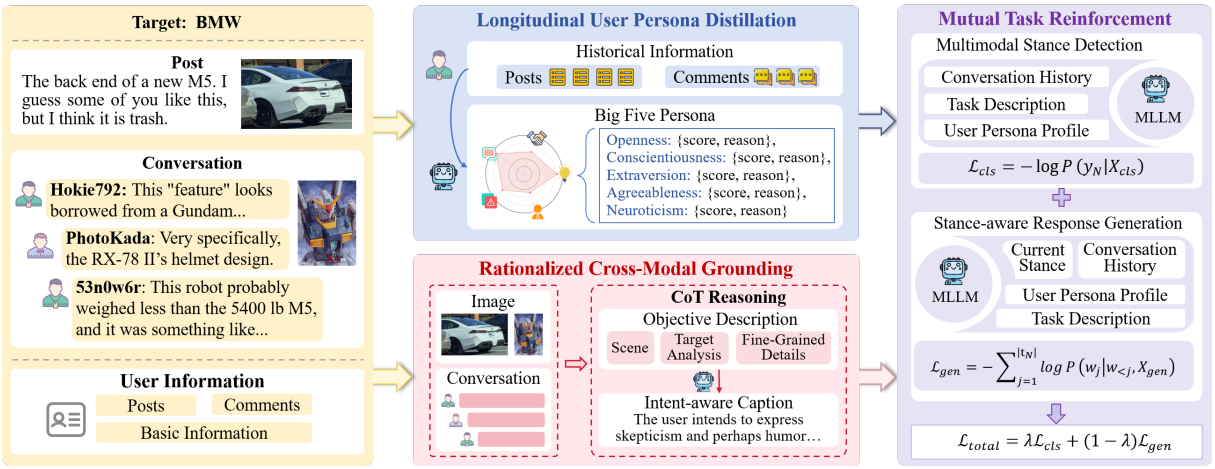

Figure 3: Overview of our PRISM framework.

senior researchers audited a randomly selected subset of the annotated data to verify accuracy and resolve borderline cases, providing an additional layer of quality assurance.

4 Our PRISM Approach

This section presents a detailed description of our proposed PRISM framework. Given a conversational thread about a global target T_g, initiated by a post $P = (t_0, v_0)$, where t_0 and v_0 are the initial text and image. The post is followed by a sequence of comments $C = \{c_1, ..., c_N\}$, where each comment c_i is a tuple (u_i, t_i, v_i) representing the user, their utterance, and an optional image. Crucially, the input also includes the multimodal historical activity H_{u_i} for each user u_i involved in the conversation, comprising their past posts and comments. Our objective is to predict the stance label y_N of the final comment c_N with respect to T_g.

As illustrated in Figure 3, PRISM comprises three key components: 1) **Longitudinal User Persona Distillation** infers each user's Big Five-based personality representation from historical posts and comments; 2) **Rationalized Cross-Modal Grounding** aligns multimodal cues with conversational context using CoT reasoning via an MLLM to generate intent-aware captions; 3) **Mutual Task Reinforcement** jointly optimizes stance detection and stance-aware response generation under a multitask objective.

4.1 Longitudinal User Persona Distillation

Stance is an externalized reflection of a user's underlying personal tendencies. Thus, each user is modeled as an entity whose stance expression is shaped by stable personality traits. For this purpose, we adopt the empirically validated Big Five (OCEAN) personality model (Jiang et al., 2024), which characterizes personality along five fundamental dimensions: Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism. To infer these traits for user u_N, we first aggregate their complete history H_{u_N}, which comprises all historical posts and comments.

Motivated by prior evidence that LLMs can provide reasonably consistent and valid personality assessments (Serapio-García et al., 2023), the distillation process leverages the full power of the MLLM $\mathcal{M}$ to reason over long-form, raw multimodal data, allowing it to capture the nuances of the user. To guide the inference, we design a detailed system instruction, $\mathcal{I}_{\text{ocean}}$. This prompt meticulously defines each of the five personality traits and specifies the output format: a numerical score from 1 to 5 for each trait, representing qualitative levels from very low to very high. The MLLM then generates a structured persona representation, p_{u_N}, composed of five numerical ratings. This process is formally expressed as:

$$\begin{aligned} p_{u_N} &= \{r^{(O)}, r^{(C)}, r^{(E)}, r^{(A)}, r^{(N)}\}_{u_N} \\ &= \mathcal{M}(\mathcal{I}_{ocean}, H_{u_N}) \end{aligned} \tag{1}$$

where each rating $r \in \{1, 2, 3, 4, 5\}$. The resulting profile p_{u_N} is a concise, structured text string that serves as a crucial, personalized context vector. This profile is explicitly injected into the prompt for the final stance detection task, ensuring the model's prediction is conditioned on the stable identity of the user in question. The prompt template can be found in Table 9 of Appendix E.

4.2 Rationalized Cross-Modal Grounding

Images in online conversations are not decorations but communicative acts whose pragmatic meaning

is shaped by the conversational context. To decipher this context-dependent rhetorical intent, we propose the Rationalized Cross-Modal Grounding (RCMG) module to generate an intent-aware caption for each image in the conversation. Inspired by Chain-of-Thought (Wei et al., 2022), this module employs a two-stage reasoning process to bridge the semantic-pragmatic gap. The complete prompt template can be found in Table 10 of Appendix E.

Objective Description. The MLLM $\mathcal{M}$ first performs a context-free analysis of the image v_i, producing a factual, objective description x_i^{obj} that grounds reasoning in the literal visual content. This step is guided by a descriptive instruction I_{desc}:

$$x_i^{obj} = \mathcal{M}(I_{desc}, v_i) \tag{2}$$

Intent-aware Interpretation. The objective description x_i^{obj} serves as an intermediate reasoning step to support the more complex task of intent inference. At this stage, the MLLM is prompted to infer the speaker's underlying motivation. We construct a prompt with a specific instruction $\mathcal{I}_{int}$, the image v_i, its objective description x_i^{obj}, and the textual content of the current conversation thread C_{text}. The model then generates the final intent-aware caption $\hat{x}_i$:

$$\hat{x}_i = \mathcal{M}(I_{int}, v_i, x_i^{obj}, C_{text}) \tag{3}$$

The resulting caption $\hat{x}_i$ captures the rhetorical function of the image within the conversation.

4.3 Mutual Task Reinforcement

To cultivate a comprehensive, user-centric understanding of stance, PRISM employs a multitask learning paradigm that jointly optimizes two inherently correlated tasks during the training phase. The core intuition is that by training the model to generate a user's response as an auxiliary task, it derives a deeper understanding of the pragmatic and user-specific cues essential for the primary task of stance prediction. This mutual reinforcement serves as a robust regularizer, enhancing the stability and prediction accuracy.

Stance Detection (Primary Task). The main objective is to predict the stance label y_N of the final comment c_N. To accomplish this, the MLLM is provided with a comprehensive set of conditioning information including the global target T_g, the persona representation of the final user p_{u_N}, the multimodal conversational comment C, and the set of all intent-aware captions $\mathcal{X}$ generated by the RCMG module for every image in the conversation. This combination of raw multimodal data and explicit textual reasoning allows the model to integrate these complementary signals. Formally, we denote the full conditioning set as:

$$X_{cls} = \{T_g, p_{u_N}, C, \mathcal{X}\} \tag{4}$$

The classification loss is the negative log-probability of the ground-truth label y_N.

$$\mathcal{L}_{cls} = -\log P(y_N | X_{cls}) \tag{5}$$

Stance-aware Response Generation (Auxiliary Task). The auxiliary task involves generating the text utterance of the final comment c_N. This task is conditioned on all preceding context, including the original multimodal conversational comments up to the penultimate turn $C_{<N} = \{c_1, ..., c_{N-1}\}$ and the corresponding set of intent-aware captions $\mathcal{X}_{<N}$ for all images appearing in the preceding turns. The conditioning information also includes the final user's persona p_{u_N} and their ground-truth stance y_N. We denote this conditioning set as:

$$X_{gen} = \{T_g, y_N, p_{u_N}, C_{<N}, \mathcal{X}_{<N}\} \tag{6}$$

The generation loss is the standard negative log-likelihood over the tokens w_j of the utterance t_N:

$$\mathcal{L}_{gen} = -\sum_{j=1}^{|t_N|} \log P(w_j | w_{<j}, X_{gen}) \tag{7}$$

Joint Optimization. The final training objective is a weighted sum of the two losses:

$$\mathcal{L}_{total} = \lambda \mathcal{L}_{cls} + (1 - \lambda) \mathcal{L}_{gen} \tag{8}$$

This mutual reinforcement ensures PRISM learns a holistic model of stance expression.

5 Experiment

5.1 Experimental Setup

Implementation details. Our framework is built on Qwen2.5-VL-7B (Bai et al., 2025) and fine-tuned for 5 epochs using our multitask objective, with the loss-weighting hyperparameter λ set to 0.7. A sensitivity analysis of λ is provided in Appendix C. We use the AdamW optimizer (Adam et al., 2014), with an initial learning rate of 1×10^{-5}, a cosine decay schedule, and a global batch size of 16. All experiments were run on a server with four

Table 2: In-target main results (%), where the methods based on LLMs involve testing through direct questioning. Best and second-best scores are in **bold** and underline, respectively. * indicates our PRISM model significantly outperforms baselines ($p < 0.05$). The dashed line separates fine-tuned from non-fine-tuned methods. w/ denotes the backbone. Ag, Fa, Avg represent F1-against, F1-favor, and F1-avg.

Modality	Method	Trump			Biden			Tesla			BMW			Costco			Bitcoin			Overall
		Ag	Fa	Avg	Ag	Fa	Avg	Ag	Fa	Avg	Ag	Fa	Avg	Ag	Fa	Avg	Ag	Fa	Avg	
Text-only	BERT	65.94	11.57	38.75	65.31	65.31	54.46	56.07	39.61	47.85	64.22	53.66	58.94	51.83	41.94	46.89	36.96	63.43	50.19	49.51
	RoBERTa	66.81	24.19	45.51	64.37	64.37	58.39	57.59	47.62	52.61	64.45	54.07	59.26	62.14	54.94	58.54	49.24	68.89	59.07	55.56
	PoliBERTweet	60.00	30.00	45.00	58.70	57.58	58.14	41.67	34.69	38.18	60.00	54.28	57.14	57.56	51.41	54.48	28.07	58.97	43.52	49.41
	LLaMA2	24.32	10.26	17.29	23.81	17.50	20.65	32.62	32.50	32.56	35.74	33.45	34.60	30.73	38.76	34.75	26.60	42.74	34.67	29.09
	GPT-3.5	62.27	27.59	44.93	51.14	58.80	54.97	63.92	45.18	54.55	60.73	50.22	55.47	62.03	52.31	57.17	47.97	65.44	56.71	53.97
	GPT-4	67.52	38.71	53.11	56.28	67.34	61.81	66.34	55.19	60.77	67.76	61.32	64.54	63.35	60.95	62.15	56.44	67.73	62.08	60.74
Multimodal	BERT+ViT	63.98	25.85	44.91	49.33	61.45	55.39	56.46	40.27	48.37	65.00	55.56	60.28	62.84	51.90	57.37	34.67	67.37	51.02	52.89
	TMPT	66.02	27.91	46.97	50.69	62.07	56.38	53.08	34.97	44.02	61.30	58.38	59.84	60.68	52.39	56.54	44.07	66.31	55.18	53.16
	Qwen-VL	62.50	24.33	43.42	49.49	54.63	52.06	50.44	44.23	47.34	53.87	42.64	48.26	62.60	54.82	58.71	50.45	69.60	60.03	51.64
	LLaVa	53.18	19.94	36.56	30.23	46.15	38.19	43.09	31.95	37.52	40.30	26.89	33.60	44.92	31.39	38.16	22.88	55.13	39.00	37.17
	MiMo	59.08	21.05	40.07	40.63	48.84	44.73	49.53	37.91	43.72	52.60	30.65	41.62	54.76	38.06	46.41	39.02	57.22	48.12	44.11
	GPT4-1	69.94	47.74	58.84	55.49	66.48	60.99	74.14	61.37	67.75	70.20	70.52	70.36	**74.88**	**72.40**	**73.64**	58.94	72.82	65.88	66.24
PRISM (ours)		73.84	48.05	60.94	58.82	**73.22***	**66.02***	72.06	**63.64***	67.85	75.00	67.61	71.30	72.44	65.72	69.08	72.00	79.45	75.72	68.49
Backbones	- w/ LLaVA	70.47	37.42	53.94	40.00	63.00	51.50	67.16	49.79	58.47	67.63	68.44	68.04	69.93	63.60	66.77	60.91	73.47	67.19	60.99
	- w/ MiMo	**74.40***	**53.01***	**63.71***	**60.61***	69.96	65.28	**75.12***	61.22	**68.17***	**75.80***	**76.86***	**76.33***	71.72	68.36	70.04	**73.22***	**81.53***	**77.38***	**70.15***

NVIDIA A100 (40 GB) GPUs using PyTorch and the Hugging Face Transformers library.

Evaluation metrics. We adopt the average F1 score (F1-avg) as the primary evaluation metric, following (et. al, 2024) and (Niu et al., 2025). F1-avg is computed as the mean of F1 scores for the against and favor stances, denoted as F1-against and F1-favor.

5.2 Baseline Methods

We conduct extensive experiments with a range of existing methods, categorized into text-only and multimodal approaches. All baseline models are fine-tuned on our specialized training data, with the exception of the GPT series, which is evaluated in a zero-shot setting. ***Text-only baselines***. We evaluate the following methods: **BERT** (Devlin et al., 2019), **RoBERTa** (Liu et al., 2019), **PoliBERTweet** (Kawintiranon and Singh, 2022), **LLaMA2** (Touvron et al., 2023), **GPT-3.5** and **GPT-4**. ***Multimodal baselines***. We include: **BERT-ViT** (Liang et al., 2024), **TMPT** (Liang et al., 2024), **LLaVA** (Liu et al., 2024), **Qwen-VL** (Bai et al., 2025), **MiMo** (Xiaomi et al., 2025) and **GPT4-1**. For more details, please refer to Appendix B.

5.3 Main Results

In-Target Stance Detection. We report the experimental results on the U-MStance dataset under the in-target setup, where the training and testing sets share identical targets. As shown in Table 2, our proposed PRISM framework achieves the highest overall performance with an F1-avg of 68.49%, which substantially outperforms multimodal GPT4-1 with 66.24% and the text-only GPT-4 with 60.74%. This improvement highlights PRISM's effectiveness and efficiency in capturing multimodal stance representations even with a relatively lightweight backbone. The results also confirm the importance of visual signals for stance understanding, as evidenced by consistent gains from multimodal models such as BERT-ViT compared to BERT, and GPT4-1 compared to GPT-4. By incorporating rationalized cross-modal grounding, PRISM effectively aligns textual and visual semantics, producing more accurate and context-aware stance predictions. Interestingly, the large models, such as LLAVA and MiMo, generally underperform compared to smaller, specialized models like TMPT. Despite their strong general language understanding, these larger models may be less effective at capturing fine-grained, conversation-specific, and multimodal cues that are critical for stance detection. In contrast, our PRISM leverages user persona and multimodal modeling to effectively capture nuanced cues, enabling its LLM backbone to achieve superior performance.

Cross-Target Stance Detection. We further evaluate generalization under the cross-target setting, where targets in the test set are unseen during training, as summarized in Table 3. Specifically, we conduct cross-target evaluations between pairs such as Trump-Biden, Tesla-BMW, and Costco-Bitcoin, training on one and testing on the other. This scenario is more challenging, requiring the model to transfer stance understanding across targets with differing semantics and discourse style.

Table 3: Comparison of different models for cross-target stance detection, where the methods based on LLMs involve testing through direct questioning. Best and second-best scores are in **bold** and underline, respectively. * indicates our PRISM model significantly outperforms baselines ($p < 0.05$). The dashed line separates fine-tuned from non-fine-tuned methods. w/ denotes the backbone. Ag, Fa, Avg represent F1-against, F1-favor, and F1-avg.

Modality	Method	Trump			Biden			Tesla			BMW			Costco			Bitcoin			Overall
		Ag	Fa	Avg	Ag	Fa	Avg	Ag	Fa	Avg	Ag	Fa	Avg	Ag	Fa	Avg	Ag	Fa	Avg	
Text-only	BERT	1.72	19.76	10.74	23.75	25.18	24.27	26.55	16.21	21.38	39.46	36.69	34.02	38.06	16.35	28.99	21.27	29.30	25.28	24.11
	RoBERTa	14.69	25.94	20.32	36.72	40.15	38.43	35.19	21.39	28.29	53.04	35.78	47.15	2.89	5.66	6.44	35.08	29.64	32.36	28.83
	PoliBERTweet	42.85	3.84	23.35	33.96	43.64	38.80	35.46	30.93	33.19	48.49	31.03	39.76	9.71	20.00	14.85	25.00	36.36	30.68	30.11
Multimodal	BERT+ViT	6.83	23.91	15.37	31.21	28.69	29.94	23.58	16.00	19.79	44.50	32.71	38.60	7.58	23.77	15.68	27.11	21.82	28.19	24.60
	TMPT	7.90	25.12	16.52	31.75	29.66	30.75	13.98	7.41	10.69	47.39	19.86	33.62	11.63	15.31	13.47	30.61	38.68	34.65	23.28
PRISM (ours)		59.50	28.43	43.97	**48.54***	61.59	55.06	62.30	**56.89***	59.60	**66.06***	51.02	58.54	45.16	**57.75***	51.45	**60.50***	67.70	64.10	55.45
Backbones	- w/ LLaVA	53.31	22.35	37.83	35.24	59.63	47.43	59.39	49.81	54.60	58.52	**58.82***	58.67	46.63	43.26	44.94	56.07	61.02	58.55	50.34
	- w/ MiMo	**60.73***	**34.39***	**47.56***	43.43	**67.03***	**55.23***	**67.90***	56.37	**62.14***	64.47	53.63	**59.05***	**47.47***	56.92	**52.20***	60.26	**68.90***	**64.58***	**56.79***

We can observe that most smaller models suffer a notable degradation in performance, revealing their limited capacity to generalize beyond the specific targets encountered during training. Remarkably, our proposed PRISM framework maintains a relatively high and stable performance across all cross-target evaluations, demonstrating superior robustness and adaptability. This advantage stems from two components: the integration of user persona modeling, which allows the model to capture consistent user-level stance tendencies across topics, and the mutual task reinforcement mechanism, which jointly optimizes stance prediction and rationalized cross-modal grounding. Together, these components enable PRISM to establish a more target-invariant representation of stance, thereby enhancing its generalization capability across heterogeneous contexts.

5.4 Impact of Different Backbone

We further assess the adaptability of PRISM by integrating it with different MLLM backbones under both the in-target and cross-target settings, as illustrated in Table 2 and Table 3. Specifically, we replace the original Qwen-VL backbone with LLaVA and MiMo to examine how backbone choice affects performance. Across both evaluation setups, MiMo provides consistent performance gains across both evaluation setups. This finding demonstrates not only the effectiveness but also the generalizability of our framework, showing that PRISM can seamlessly adapt to stronger multimodal backbones while preserving its core advantages. The gains from MiMo likely stem from its enhanced multimodal reasoning capabilities and more robust cross-domain representation learning. These properties align well with PRISM's design, enabling more precise modeling of nuanced stance cues and improving the transferability of stance understanding across heterogeneous targets.

Table 4: Experimental results (%) of the ablation study. w/o means without.

	Trump	Biden	Tesla	BMW	Costco	Bitcoin
In-Target						
PRISM	**60.94**	**66.02**	**67.85**	**71.30**	**69.08**	**75.72**
-w/o Persona	49.56	60.38	61.67	70.02	64.85	72.92
-w/o Intent	50.81	57.30	56.45	65.34	61.60	72.59
-w/o Mutual	59.64	64.34	67.07	68.97	68.89	73.34
Cross-Target						
PRISM	**43.97**	**55.06**	**59.60**	**58.54**	**51.45**	**64.10**
-w/o Persona	40.96	47.16	56.42	55.95	50.13	60.73
-w/o Intent	40.99	45.19	41.90	48.30	47.84	53.66
-w/o Mutual	43.63	52.16	58.80	54.01	49.88	63.91

5.5 Ablation Study

To evaluate the contribution of each component in our PRISM framework, we conduct a comprehensive ablation study with the results presented in Table 4. Replacing the persona representation p_{u_N} with the user's original raw information (**w/o Persona**) consistently drops performance across all targets. This underscores the critical role of persona distillation in filtering inherent noise and extracting task-relevant characteristics, enabling PRISM to more precisely interpret individualized, user-centric stance expressions.

Substituting intent-aware captions $\hat{x}$ with raw visual features (**w/o Intent**) results in a more substantial degradation, highlighting the crucial role of rationalized cross-modal grounding. By explicitly aligning the image pragmatic intent with textual context, this component enables the PRISM to correctly infer subtle or implicit visual cues essential for multimodal stance detection. Lastly, discarding the mutual task reinforcement mechanism and reverting to a single-task stance detection objective (**w/o Mutual**) also leads to a noticeable decline in performance. This confirms that jointly optimizing stance detection with auxiliary generation task encourages beneficial interactions

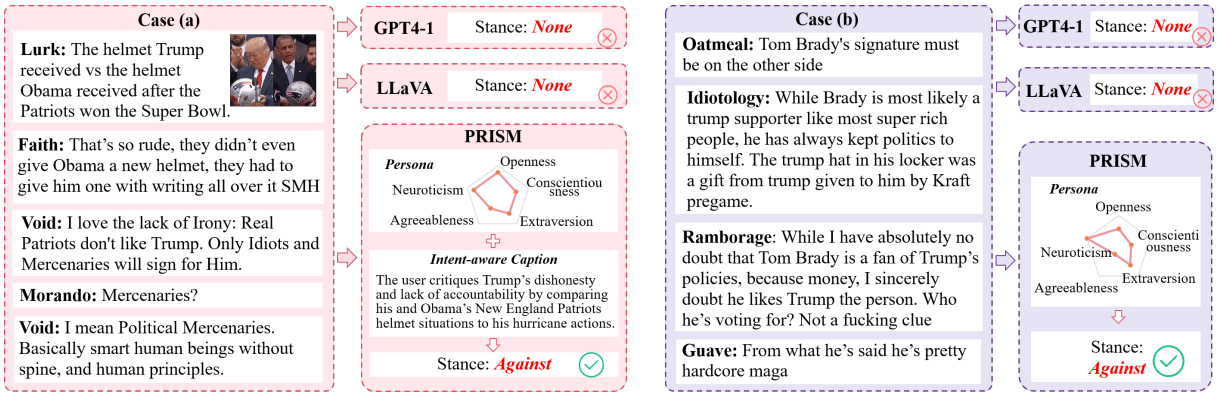

Figure 4: Case study comparing PRISM with baseline models.

between multimodal understanding and stance reasoning, ultimately producing more robust and transferable representations. When all three components are removed, the resulting variant reduces to the backbone-only baseline reported in Table 2.

5.6 Case Study

To visually demonstrate the superiority of our method and the inherent challenges of the task, Figure 4 provides a case study. In Case (a), both GPT4-1 and LLaVA fail to detect the sarcastic implication conveyed by joint multimodal cues, incorrectly assigning a neutral stance due to their reliance on surface-level information. In contrast, our PRISM accurately predicts an *Against* stance. This improvement stems from its rationalized cross-modal grounding, which captures the user's implicit criticism expressed through the comparison embedded in the image, as well as its persona-aware interpretation of the user's communicative tendencies.

Meanwhile, Case (b) presents a more subtle scenario, where sarcasm is intertwined with the user's disposition. Again, GPT4-1 and LLaVA misclassify the stance, likely due to their lack of user-level modeling and reliance solely on the current conversational snippet. PRISM, however, correctly identifies the *Against* stance by leveraging its persona module to interpret longitudinal behavioral patterns. Specifically, the persona visualization highlights a high Neuroticism score, reflecting an emotionally expressive and critical style. Incorporating such personality cues enables PRISM to infer the user's consistent critical orientation even when sarcasm is subtle.

5.7 Impact of Conversation Depth

We further evaluate how PRISM's stance detection performance is influenced by varying conversational depths. As shown in Figure 5, performance shows a gradual decline as depth increases, which is expected since deeper conversations involve more intricate discourse structures, longer-range dependencies, and greater contextual ambiguity, making the final stance harder to infer. Notably, for the targets *Costco* and *Bitcoin*, PRISM maintains stable performance across depths. This robustness is attributed to the rationalized cross-modal grounding module, whose intent-aware image captions align visual and textual cues and preserve salient multimodal information, thereby mitigating challenges from longer conversational contexts.

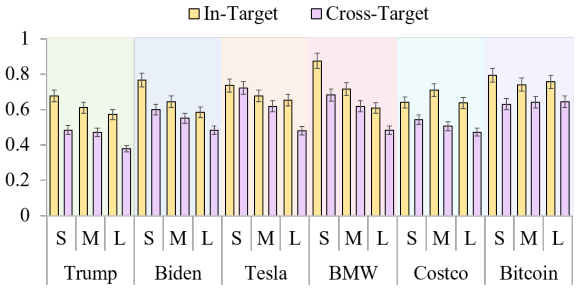

Figure 5: Performance of different depths. Conversational depth: S (1-2 turns), M (3-4 turns), L (≥5 turns).

6 Conclusion

This paper presents two critical limitations in MCSD, including pseudo-multimodality, where visual cues are limited to source posts, and user homogeneity, which neglects individual differences. To address these, we propose U-MStance, the first user-centric MCSD dataset with multimodal and user information. Furthermore, we design PRISM, a framework that distills stable user personas, grounds image intent, and reinforces stance understanding through multitask learning. Extensive experiments on U-MStance demonstrate that PRISM consistently outperforms strong baselines and exhibits robust generalization.

Limitations

Despite the significant improvements achieved by our model, certain limitations remain. First, while U-MStance spans diverse domains including politics, automotive, consumer goods, and cryptocurrency, it lacks coverage of highly specialized or professional sectors, such as specific legal statutes or cutting-edge scientific controversies. Second, as illustrated in Figure 5, model performance exhibits a declining trend as conversational depth increases. In extremely deep or multi-branch discussion threads, the model still encounters challenges in perfectly capturing subtle logical transitions and long-range contextual drifts.

References

Kingma DP Ba J Adam and 1 others. 2014. A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 1412(6).

Emily Allaway and Kathleen McKeown. 2020a. Zero-shot stance detection: A dataset and model using generalized topic representations. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8913–8931.

Emily Allaway and Kathleen McKeown. 2020b. Zero-shot stance detection: A dataset and model using generalized topic representations. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8913–8931.

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*.

Erik Cambria, Soujanya Poria, Devamanyu Hazarika, and Kenneth Kwok. 2018. Senticnet 5: Discovering conceptual primitives for sentiment analysis by means of context embeddings. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32.

Zebang Cheng, Zhi-Qi Cheng, Jun-Yan He, Kai Wang, Yuxiang Lin, Zheng Lian, Xiaojiang Peng, and Alexander Hauptmann. 2024a. Emotion-llama: Multimodal emotion recognition and reasoning with instruction tuning. *Advances in Neural Information Processing Systems*, 37:110805–110853.

Zebang Cheng, Fuqiang Niu, Yuxiang Lin, Zhi-Qi Cheng, Xiaojiang Peng, and Bowen Zhang. 2024b. Mips at semeval-2024 task 3: Multimodal emotion-cause pair extraction in conversations with multimodal language models. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 667–674.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

Costanza Conforti, Jakob Berndt, Mohammad Taher Pilehvar, Chryssi Giannitsarou, Flavio Toxvaerd, and Nigel Collier. 2020. Will-they-won't-they: A very large dataset for stance detection on twitter. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1715–1724.

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.

Daijun Ding, Rong Chen, Liwen Jing, Bowen Zhang, Xu Huang, Li Dong, Xiaowen Zhao, and Ge Song. 2024a. Cross-target stance detection by exploiting target analytical perspectives. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 10651–10655. IEEE.

Daijun Ding, Li Dong, Zhichao Huang, Guangning Xu, Xu Huang, Bo Liu, Liwen Jing, and Bowen Zhang. 2024b. Edda: An encoder-decoder data augmentation framework for zero-shot stance detection. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 5484–5494.

Niu et. al. 2024. Multimodal multi-turn conversation stance detection: A challenge dataset and effective model. In *ACM MM*.

Kyle Glandt, Sarthak Khanal, Yingjie Li, Doina Caragea, and Cornelia Caragea. 2021. Stance detection in covid-19 tweets. In *Proceedings of the 59th annual meeting of the association for computational linguistics and the 11th international joint conference on natural language processing (long papers)*, volume 1.

Hang Jiang, Xiajie Zhang, Xubo Cao, Cynthia Breazeal, Deb Roy, and Jad Kabbara. 2024. Personallm: Investigating the ability of large language models to express personality traits. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3605–3627.

Kornraphop Kawintiranon and Lisa Singh. 2022. Polibertweet: a pre-trained language model for analyzing political content on twitter. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 7360–7367.

Xiaochong Lan, Chen Gao, Depeng Jin, and Yong Li. 2024. Stance detection with collaborative role-infused llm-based agents. In *Proceedings of the international AAAI conference on web and social media*, volume 18, pages 891–903.

Ang Li, Bin Liang, Jingqian Zhao, Bowen Zhang, Min Yang, and Ruifeng Xu. 2023a. Stance detection on social media with background knowledge. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pages 15703–15717.

Ang Li, Jingqian Zhao, Bin Liang, Lin Gui, Hui Wang, Xi Zeng, Kam-Fai Wong, and Ruifeng Xu. 2024. Mitigating biases of large language models in stance detection with calibration. *CoRR*.

Yingjie Li and Cornelia Caragea. 2021. A multi-task learning framework for multi-target stance detection. In *Findings of the association for computational linguistics: ACL-IJCNLP 2021*, pages 2320–2326.

Yingjie Li, Tiberiu Sosea, Aditya Sawant, Ajith Jayaraman Nair, Diana Inkpen, and Cornelia Caragea. 2021. P-stance: A large dataset for stance detection in political domain. In *Findings of the association for computational linguistics: ACL-IJCNLP 2021*, pages 2355–2365.

Yupeng Li, Haorui He, Shaonan Wang, Francis CM Lau, and Yunya Song. 2023b. Improved target-specific stance detection on social media platforms by delving into conversation threads. *IEEE Transactions on Computational Social Systems*, 10(6):3031–3042.

Yupeng Li, Haorui He, Shaonan Wang, Francis CM Lau, and Yunya Song. 2023c. Improved target-specific stance detection on social media platforms by delving into conversation threads. *IEEE Transactions on Computational Social Systems*, 10(6):3031–3042.

Yupeng Li, Dacheng Wen, Haorui He, Jianxiong Guo, Xuan Ning, and Francis CM Lau. 2023d. Contextual target-specific stance detection on twitter: Dataset and method. In *2023 IEEE International Conference on Data Mining (ICDM)*, pages 359–367. IEEE.

Bin Liang, Zixiao Chen, Lin Gui, Yulan He, Min Yang, and Ruifeng Xu. 2022. Zero-shot stance detection via contrastive learning. In *Proceedings of the ACM web conference 2022*, pages 2738–2747.

Bin Liang, Yonghao Fu, Lin Gui, Min Yang, Jiachen Du, Yulan He, and Ruifeng Xu. 2021. Target-adaptive graph for cross-target stance detection. In *Proceedings of the web conference 2021*, pages 3453–3464.

Bin Liang, Ang Li, Jingqian Zhao, Lin Gui, Min Yang, Yue Yu, Kam-Fai Wong, and Ruifeng Xu. 2024. Multi-modal stance detection: New datasets and model. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 12373–12387.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 26296–26306.

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*.

Saif Mohammad, Svetlana Kiritchenko, Parinaz Sobhani, Xiaodan Zhu, and Colin Cherry. 2016. Semeval-2016 task 6: Detecting stance in tweets. In *Proceedings of the 10th international workshop on semantic evaluation (SemEval-2016)*, pages 31–41.

Fuqiang Niu, Min Yang, Ang Li, Baoquan Zhang, Xiaojiang Peng, and Bowen Zhang. 2024. A challenge dataset and effective models for conversational stance detection. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 122–132.

Fuqiang Niu, Yi Yang, Xianghua Fu, Genan Dai, and Bowen Zhang. 2025. C-mtcsd: A chinese multi-turn conversational stance detection dataset. In *Companion Proceedings of the ACM on Web Conference 2025*, pages 769–772.

Fuqiang et. al Niu. 2025. Mt2-csd: A new dataset and multi-semantic knowledge fusion method for conversational stance detection. *arXiv preprint arXiv:2506.21053*.

Greg Serapio-García, Mustafa Safdari, Clément Crepy, Luning Sun, Stephen Fitz, Peter Romero, Marwa Abdulhai, Aleksandra Faust, and Maja Matarić. 2023. Personality traits in large language models. *arXiv preprint arXiv:2307.00184*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, and 1 others. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Ramon et. al Villa-Cox. 2020. Stance in replies and quotes (srq): A new dataset for learning stance in twitter conversations. *arXiv preprint arXiv:2006.00691*.

Bingbing Wang, Zhixin Bai, Qianlong Wang, Jingjie Lin, Min Yang, Xi Zeng, and Ruifeng Xu. 2025. More than just a conversation: A multi-agent reasoning graph knowledge distillation for conversational stance detection. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2895–2899.

Xiaolong Wang, Yile Wang, Sijie Cheng, Peng Li, and Yang Liu. 2024. Deem: Dynamic experienced expert modeling for stance detection. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 4530–4541.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Penghui Wei and Wenji Mao. 2019. Modeling transferable topics for cross-target stance detection. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1173–1176.

Maxwell Weinzierl and Sanda Harabagiu. 2023. Identification of multimodal stance towards frames of communication. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pages 12597–12609.

LLM Xiaomi, Bingquan Xia, Bowen Shen, Dawei Zhu, Di Zhang, Gang Wang, Hailin Zhang, Huaqiu Liu, Jiebao Xiao, Jinhao Dong, and 1 others. 2025. Mimo: Unlocking the reasoning potential of language model–from pretraining to posttraining. *arXiv preprint arXiv:2505.07608*.

Bowen Zhang, Daijun Ding, Zhichao Huang, Ang Li, Yangyang Li, Baoquan Zhang, and Hu Huang. 2024. Knowledge-augmented interpretable network for zero-shot stance detection on social media. *IEEE Transactions on Computational Social Systems*.

Bowen Zhang, Daijun Ding, Liwen Jing, Genan Dai, and Nan Yin. 2022. How would stance detection techniques evolve after the launch of chatgpt? *arXiv preprint arXiv:2212.14548*.

Bowen Zhang, Xianghua Fu, Daijun Ding, Hu Huang, Genan Dai, Nan Yin, Yangyang Li, and Liwen Jing. 2023. Investigating chain-of-thought with chatgpt for stance detection on social media. *arXiv preprint arXiv:2304.03087*.

Qinglin Zhu, Bin Liang, Jingyi Sun, Jiachen Du, Lanjun Zhou, and Ruifeng Xu. 2022. Enhancing zero-shot stance detection via targeted background knowledge. In *Proceedings of the 45th international ACM SIGIR conference on research and development in information retrieval*, pages 2070–2075.

A Detailed Dataset Statistics

Figure 6 illustrates the distribution of instances across different labels. Table 5 illustrates the distribution of instances across different conversational depths, including the number of users and images for each target. Within our collected dataset, *36%* of the posts contain a single image, while *63%* include two or more images, which means visual information is also present within the comments. In general, our dataset features more extensive conversational depth compared to existing conversational stance detection datasets, and achieves 100% vision-related coverage.

For model development, the dataset is partitioned into training, validation, and test sets using an 8:1:1 ratio. To reduce potential leakage, our construction applies three controls. (1) User-profile information is temporally bounded: only posts and comments created before the target sample timestamp are retained, so no future user evidence is introduced. (2) Dataset splitting is performed at the post level rather than at the instance level, which prevents shared dialogue context across train/dev/test splits. (3) After anonymization and sample linking to the current speaker, we additionally checked cross-split user overlap and removed the small number of remaining cases. These steps jointly reduce future leakage, context leakage, and user-identity leakage.

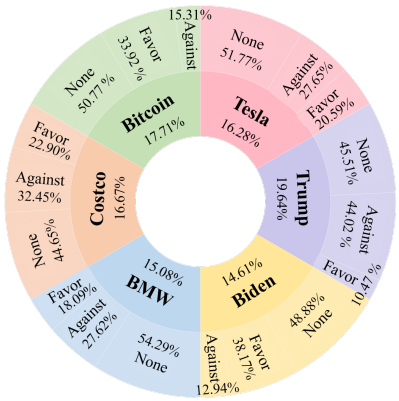

Figure 6: The distribution of stance categories in our U-MStance.

The results show clear sensitivity to task weighting, which is consistent with the fact that the classification objective and the auxiliary generation objective impose different optimization preferences in multitask training. Extremely small or extremely large values lead to weaker overall performance, whereas a more balanced weighting yields the strongest result. In the current setting, performance peaks at $\lambda = 0.7$ with an overall F1-avg of 66.41.

B Model Details

Table 7 summarizes all baseline models used in our experiments, while Table 8 provides their detailed configurations, where Conversation denotes the Conversation History, Image refers to the visual modality within the conversation, and Profile represents the Reddit user profile associated with each sample. All open-source models were fine-tuned on our dataset.

Table 5: Comparison of Depth and Interaction Data Across Different Entities

	Depth	Avg. WC	Trump	Biden	Tesla	BMW	Costco	Bitcoin	Total
User	–	–	5087	3371	3896	3956	4519	4123	24952
Image	–	–	47	40	55	102	20	21	285
Post	1	17.96	16	27	14	19	8	13	97 (0.24%)
Comment	2	27.38	927	1011	723	741	768	1186	5356 (13.39%)
	3	23.71	2306	2096	1982	1816	2288	2647	13135 (32.84%)
	4	27.65	1645	1189	1402	1176	1470	1292	8174 (20.43%)
	5	29.99	1076	694	928	821	1001	846	5366 (13.41%)
	6	31.52	795	427	621	637	541	517	3538 (8.84%)
	7	35.98	529	231	430	385	316	304	2195 (5.49%)
	8	39.59	333	106	264	290	168	173	1334 (3.33%)
	9	45.02	231	65	150	148	108	106	808 (2.02%)

Table 6: Sensitivity analysis of the loss-weighting coefficient λ.

λ	Trump	Biden	Tesla	BMW	Costco	Bitcoin	Overall
0.3	39.99	53.74	53.35	56.01	58.31	56.51	52.99
0.5	44.06	56.91	57.49	57.96	66.06	60.17	57.11
0.7	57.43	66.31	69.17	68.43	72.41	64.68	66.41
0.9	47.72	60.03	57.58	62.13	66.16	61.74	59.23

Table 7: Model details including parameters, access type, and their respective sources.

Model	Parameters	Access	Source
BERT	<1B	Local	Google
RoBERTa	<1B	Local	Meta
PoliBERTweet	<1B	Local	Kawintiranon et al.
LLaMA-2-7B	7B	Local	Meta
GPT-3.5	Unknown	API	OpenAI
GPT-4	Unknown	API	OpenAI
BERT-ViT	<1B	Local	Liang et al.
TMPT	<1B	Local	Liang et al.
LLaVA-1.5-7B	7B	Local	Microsoft
Qwen2.5-VL-7B	7B	Local	Alibaba
MiMo-VL-7B	7B	Local	Xiaomi
GPT4-1	Unknown	API	OpenAI

Table 8: The experimental setup for baseline models.

Model	Input			Output
	Conversation	Image	Profile	
BERT	✓	✗	✓	Stance Class
RoBERTa	✓	✗	✓	Stance Class
PoliBERTweet	✓	✗	✓	Stance Class
LLaMA-2-7B	✓	✗	✓	Stance Class
GPT-3.5	✓	✗	✓	Stance Class
GPT-4	✓	✗	✓	Stance Class
BERT-ViT	✓	✓	✓	Stance Class
TMPT	✓	✓	✓	Stance Class
LLaVA-1.5-7B	✓	✓	✓	Stance Class
Qwen2.5-VL-7B	✓	✓	✓	Stance Class
MiMo-VL-7B	✓	✓	✓	Stance Class
GPT4-1	✓	✓	✓	Stance Class

C Sensitivity Analysis of λ

In our setting, λ is selected empirically on the validation set. The sensitivity analysis results over $\lambda \in \{0.3, 0.5, 0.7, 0.9\}$ are shown in Table 6.

D Screenshot of the Annotation Tool

Figure 7 and Figure 8 illustrate the screenshots of the data annotation interface.

E Prompt Templates

To support the reproducibility and interpretability of our methodology and experimental settings, we provide the comprehensive set of prompt templates used in this work. Specifically, Table 9 details the prompts for User Persona Distillation; Table 10 presents the two-stage prompt protocol for Image Intent Extraction; Table 11 outlines the prompt employed for the baseline models; and Table 12 outlines the prompt employed in our method.

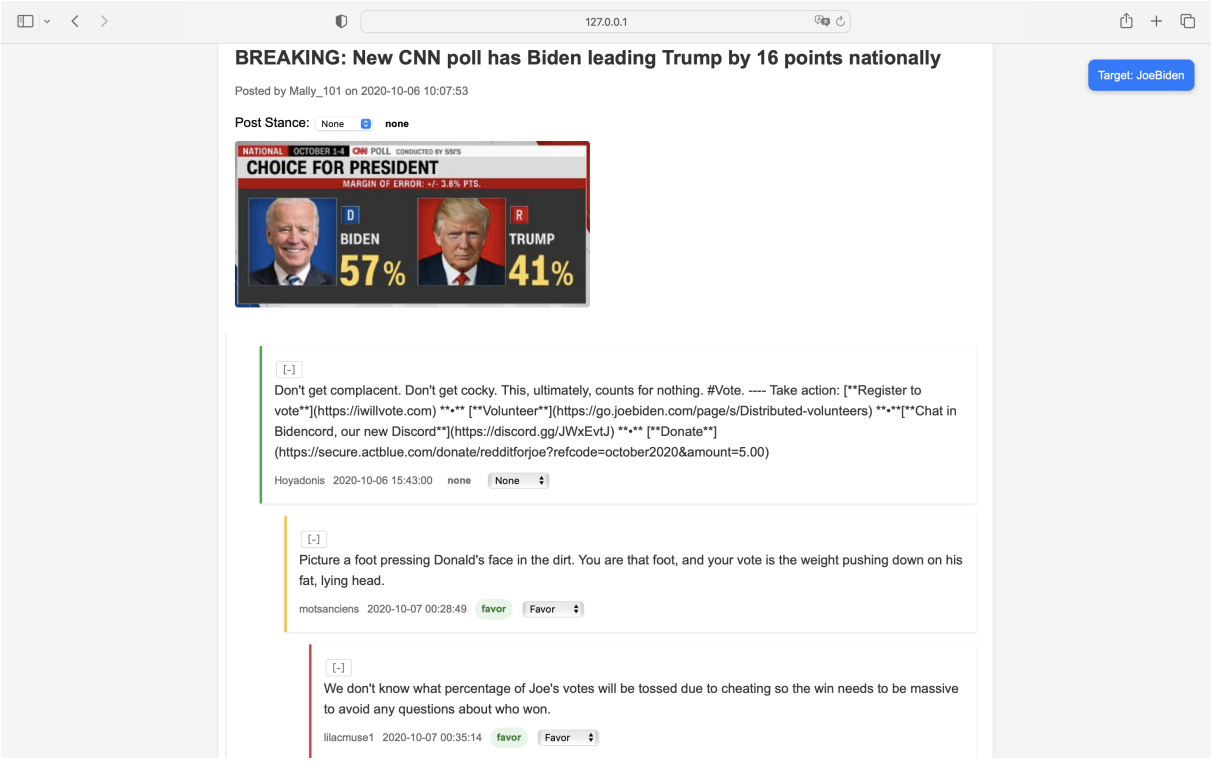

Figure 7: The Web Interface of the Data Annotation Process.

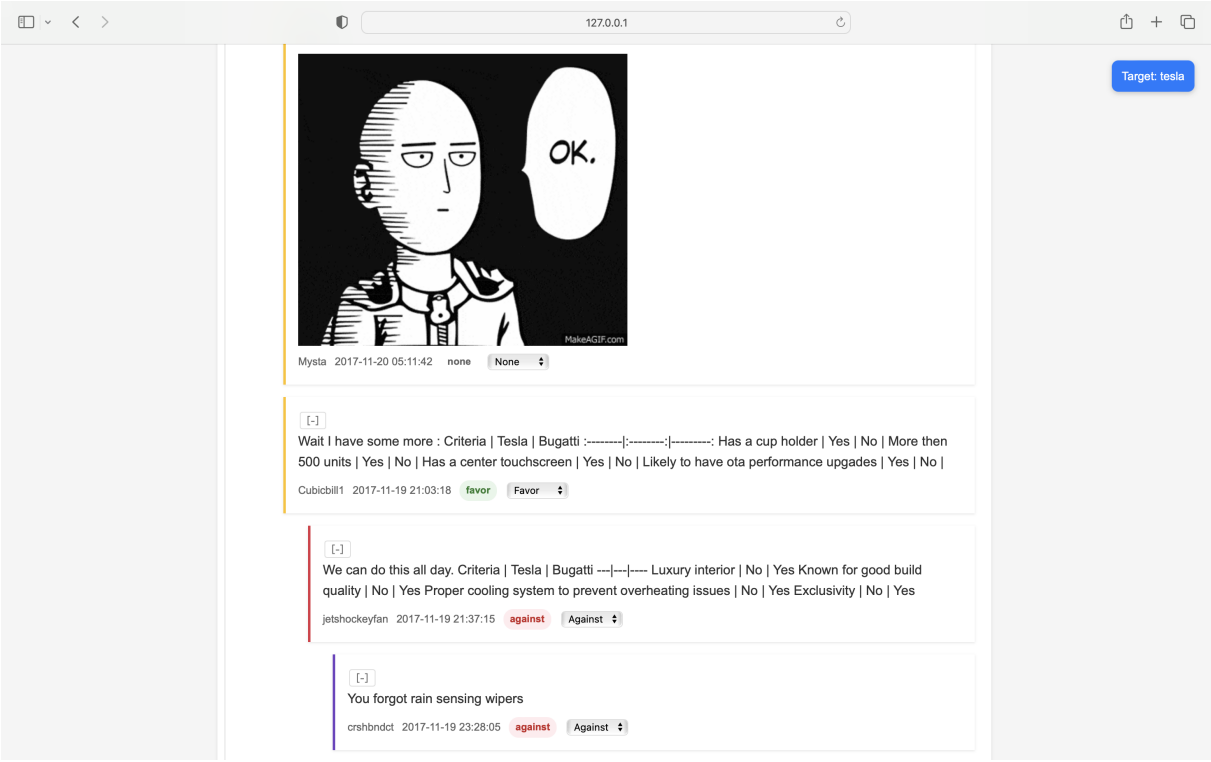

Figure 8: The Web Interface of the Data Annotation Process.

B.1 User Persona Distillation

System Prompt:
You are an expert in computational psychology and linguistic analysis. Your task is to analyze a user's Big Five personality traits (Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism) based on their historical social media activity.

User Prompt:
Task Description:
Analyze the provided user profiles, including their posts and comments. For each of the five traits, provide a qualitative description of how the user exhibits this trait and a quantitative score on a scale of 1-5 (1: Very Low, 5: Very High).

Big Five Traits Definition:
1. Openness to Experience: Curiosity, originality, and openness to new ideas.
2. Conscientiousness: Organization, responsibility, and goal-oriented behavior.
3. Extraversion: Sociability, assertiveness, and emotional expressiveness.
4. Agreeableness: Trust, altruism, kindness, and affection.
5. Neuroticism: Emotional instability, anxiety, and irritability.

Input Data:
User ID: {{user id}}
Recent Posts:
{{post list}}
Recent Comments:
{{comment list}}

Output:

Table 9: Prompt template for User Persona Distillation. The model outputs a summary of Big Five personality traits for users.

B.2 Image Intent Extraction

Step 1:

System Prompt:
You are an expert image analyzer.

User Prompt:
Task Description:
You will be given an image. Your task is to produce a detailed and structured description of the image. Please follow the required steps:
1. **Overall Scene Description**
Describe the setting, environment, and general context of the image.
2. **Key Objects and Entities**
Identify and describe the notable objects, people, or elements in the image, including their appearance, attributes, and spatial relationships.
3. **Fine-Grained Visual Details**
Highlight colors, expressions, textures, gestures, body language, or any small details that may convey meaning.
4. **Style and Emotional Tone (if applicable)**
Describe the mood, atmosphere, artistic style, or emotional impression conveyed by the image.

Input: <image>

Output Requirment:
Make sure your description is precise, neutral, and avoids assumptions unless clearly supported by visual evidence.
Directly output the description text**, starting with **'Image Description:'** followed by your concise answer.

Output:

Step 2:

System Prompt:
You are an expert in multimodal conversation analysis.

User Prompt:
Task Description:
Determine the **intention or communicative purpose** of the user who posted the image in the conversation.
Consider the social and semantic context of the conversation, as well as the content of the image.

You are given:
1. A **description of the image**, which has been generated by a visual-language model.
2. A **conversation history** between multiple participants.
- In this history, the placeholder "<image>" marks where the image was posted by the user.

If the conversation contains multiple images, carefully identify the **index or order** of the images (e.g., <image1>, <image2>, <image3>...) and determine where the **current target image** appears in the conversation before analyzing it.

Requirment:
Follow these steps carefully:
1. Review the conversation history and understand the flow of dialogue.
2. Interpret the meaning of the image based on the provided description.
3. Analyze how the image contributes to the conversation — for example, whether it expresses emotion, humor, evidence, sarcasm, agreement, mockery, or provides visual information.
4. Infer the **most likely intention** of the user when posting the image.
5. Directly output the intent text**, starting with **'Intent:'** followed by your concise answer.

Image Description:
The image captures a bustling urban scene at night, likely Times Square in New York City, characterized by its iconic billboards and vibrant advertisements. The foreground features ...

Conversation History:
User youyouxue: We just drove from San Francisco to New York City in a Tesla Model X without using a single drop of gas! <image1>
User salec1:I did a shorter version of the same thing (Miami to NYC) last month. It's all about ...
User KickAClay:I always wondered if any Tesla, after inputting your route ...
User timdorr:Yes, it does exactly this. Set your destination and it will find Superchargers along your route, tell you how long to ...
User chupacabrando:How much longer does it take ...
...

Output:

Table 10: Prompt template for Image Intent Extraction. The model outputs a detailed and structured description of the image.

B.3 Multimodal Conversational Stance Detection

System Prompt:
You are an expert linguist specializing in conversational analysis and social media stance detection.

User Prompt:
Task Description:
Analyze the provided conversation flow and determine the stance of the current speaker toward the specific Target listed below. The attached image(s) correspond to the sequential labels (e.g., <image1>, <image2>) mentioned in the text.

Target:
Tesla

Stance Classes:
['favor','against','none']

Conversation History:
User youyouxue: We just drove from San Francisco to New York City in a Tesla Model X without using a single drop of gas! <image1>
User salec1:I did a shorter version of the same thing (Miami to NYC) last month. It's all about ...
User KickAClay:I always wondered if any Tesla, after inputting your route ...
User timdorr:Yes, it does exactly this. Set your destination and it will find Superchargers along your route, tell you how long to ...
User chupacabrando:How much longer does it take ...
... ### Current Speaker's Reddit Profile:

Active Communities:
— Historical Posts —
...
— Historical Comments —
...

Output Format (JSON):
s̈tance": "favor" | "against" | "none"

Output:

Table 11: Prompt template for Multimodal Conversational Stance Detection. The model outputs a stance class for the current instance.

B.4 Multimodal Conversational Stance Detection (Our method)

System Prompt:
You are an expert linguist specializing in conversational analysis and social media stance detection.

User Prompt:
Task Description:
Analyze the provided conversation flow and determine the stance of the current speaker toward the specific Target listed below. The attached image(s) correspond to the sequential labels (e.g., <image1>, <image2>) mentioned in the text.

Target:
Tesla

Stance Classes:
['favor','against','none']

Conversation History:
User youyouxue: We just drove from San Francisco to New York City in a Tesla Model X without using a single drop of gas! <image1>
User salec1:I did a shorter version of the same thing (Miami to NYC) last month. It's all about ...
User KickAClay:I always wondered if any Tesla, after inputting your route ...
User timdorr:Yes, it does exactly this. Set your destination and it will find Superchargers along your route, tell you how long to ...
User chupacabrando:How much longer does it take ...
...

Image Intent Summary:
The Intent of <image1>: The user intends to share the image to express their confidence in Biden's election victory and to highlight the simulation results showing a high probability of Biden winning. The user likely aims to provide evidence-based information to support their stance on the election outcome.
...

Current Speaker's Big Five Personality Analysis:
- Openness (4/5 - High): Curious and open to new experiences.
- Conscientiousness (3/5 - Moderate): Organized and responsible, with moderate discipline.
- Extraversion (2/5 - Low): Sociable but not overly expressive.
- Agreeableness (3/5 - Moderate): Pragmatic and direct, sometimes skeptical.
- Neuroticism (3/5 - Moderate): Emotionally stable with occasional stress.

Output Format (JSON):
"stance": "favor" | "against" | "none"

Output:

Table 12: Prompt template for Multimodal Conversational Stance Detection in our method. The model outputs a stance class for the current instance.