# Cause-CSD: A Challenge Multimodal Conversational Stance Cause Detection Dataset and Effective Method

**Fuqiang Niu[1]*, Bowen Zhang[2]*, Junting Zhu[2], Qing Liao[3], Genan Dai[2], Hu Huang[1]†**
[1]School of Cyber Science and Technology,
University of Science and Technology of China, Hefei, China
[2]School of Artificial Intelligence, Shenzhen Technology University, Shenzhen, China
[3]School of Computer Science and Technology,
Harbin Institute of Technology, Shenzhen, China

## Abstract

Social media platforms have become critical arenas for public discourse, yet existing stance detection methods often reduce opinions to surface-level labels, overlooking the conversational evidence behind stance expressions. We introduce Conversational Stance-Cause Pair Detection (CSCPD), a new task that jointly identifies both the stance polarity and its observable contextual evidence within multi-turn conversations. To advance research in this direction, we present Cause-CSD, the first large-scale dataset for CSCPD, spanning 21,048 annotated stance-cause pairs across diverse open-domain, textual, and multimodal discussions. We further propose Stance-Cause Detection Language Model (SCD-LM), a unified language model framework that leverages explicit context reasoning and joint decoding to predict stances and their supporting causes, along with human-readable rationales. Extensive experiments demonstrate that SCD-LM achieves state-of-the-art results on both text-only and multimodal subtasks, significantly outperforming strong baselines, especially for long-range and image-grounded cause detection. Our work advances explainable stance analysis and underpins understanding of public opinion drivers in impactful online settings.

## 1 Introduction

Social media platforms have become the global epicenter for public debate, offering a real-time record of collective attitudes on contentious topics ranging from public health to elections (Glandt et al., 2021). To analyze these dynamics, Conversational Stance Detection (CSD) has emerged as a critical tool, aiming to identify the polarity (e.g., favor, against) of opinions in multi-turn discussions (Somasundaran and Wiebe, 2010; Augenstein et al., 2016; Küçük and Can, 2020). However, despite their prevalence, existing CSD approaches treat stance as an isolated, surface-level label, abstracted away from the conversational events or evidence that triggered it. This limitation leaves a critical gap: systems may classify stances with high accuracy yet remain blind to the observable conversational evidence and contextual grounding associated with stance expressions.

This lack of explicit conversational grounding severely limits the practical utility of stance analysis in high-stakes domains. In scenarios such as crisis response and policy-making, the critical question is not merely what stance a speaker expresses, but what observable contextual evidence in the conversation is associated with that stance expression. Stance labels alone fail to indicate whether the stance is grounded in factual claims, emotional reactions, or other observable contextual evidence, making it difficult to design targeted interventions such as counter-messaging or policy adjustments. Consequently, to prevent misinterpretation and enable effective decision-making, it is imperative to move beyond simple stance identification and explicitly model the observable conversational evidence behind stance expressions in discourse (Saha et al., 2024; Li et al., 2025).

Motivated by this observation, we argue that stance in conversational settings is typically grounded in specific conversational context rather than being a static attribute of an utterance (Li et al., 2023d). Such context may include the utterance itself, earlier arguments, claims, or external evidence introduced in the discussion, and in many cases spans multiple turns or modalities. Without explicitly identifying the observable contextual evidence for a stance expression, stance interpretation remains incomplete, particularly in complex or multimodal conversations (Niu et al., 2024a).

Building on this perspective, we introduce a new task: Conversational Stance-Cause Pair Detection (CSCPD). CSCPD aims to jointly identify both

*These authors contributed equally.
†Corresponding authors: huanghu@ustc.edu.cn



*Findings of the Association for Computational Linguistics: ACL 2026*, pages 38654–38666

the stance polarity expressed by a speaker and the specific conversational content that serves as the observable contextual evidence for that stance expression. In this work, we use the term "cause" in an operational sense to denote observable contextual evidence in the thread, rather than the speaker's latent beliefs, long-term motivations, or broader belief formation process. By explicitly linking stance expressions to their supporting conversational evidence, CSCPD enables a more grounded analysis of conversational dynamics and provides insights that go beyond stance labels alone.

CSCPD differs from existing work on stance detection and post-hoc explanation in several important ways. Rather than treating causes as optional rationales generated after stance prediction, CSCPD models stance and cause as inherently interdependent: a stance cannot be correctly understood without identifying the observable contextual evidence associated with it. Moreover, stance causes may be implicit, target-dependent, and distributed across long conversational contexts or multiple speakers. These characteristics introduce unique challenges, including long-range context modeling, cross-turn reasoning, and the need to ensure logical consistency between predicted stances and their supporting causes.

Despite the relevance of this problem, progress has been limited by the lack of datasets explicitly designed to support stance–cause modeling in conversations. To address this gap, we present Cause-CSD, the first large-scale dataset tailored for the CSCPD task. Cause-CSD consists of 21,048 annotated stance–cause pairs collected from diverse, open-domain conversational settings. The dataset includes two subtasks: (Subtask A) text-only conversations, where both stance and cause are inferred from dialogue, and (Subtask B) multimodal conversations, where textual and visual content jointly inform stance formation. Figure 1 illustrates examples from both subtasks, highlighting how visual evidence (e.g., an image depicting a burned vehicle) can play a decisive role in shaping stance.

To effectively address the challenges posed by CSCPD, we further propose the Stance-Cause Detection Language Model (SCD-LM), a unified framework for jointly predicting stance and its cause. SCD-LM adopts a two-stage architecture comprising Context Reasoning and Joint Decoding. In the first stage, to capture implicit semantics and long-range dependencies, a large language model is leveraged as a teacher to reason over the conversational context and identify candidate stance–cause pairs along with concise natural-language rationales. In the second stage, a lightweight decoder is fine-tuned on this rationale-augmented supervision to jointly predict stance polarity and the index of the corresponding cause within output sequence. This joint formulation mitigates error propagation commonly observed in pipeline approaches and promotes consistency between predicted stances and their supporting evidence.

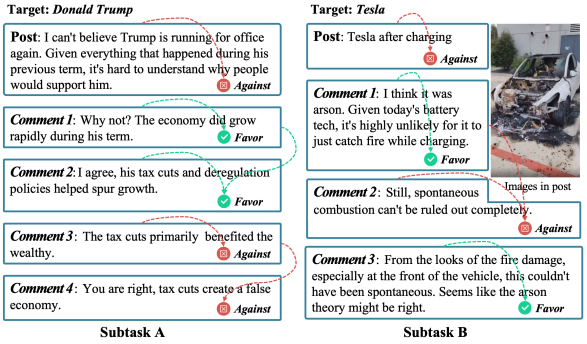


Figure 1: An Example of the CSCPD Task. Dashed arrows denote directed alignments from each stance-bearing utterance to its annotated conversational evidence in the thread. The evidence may be the utterance itself, one or more prior utterances, or multimodal context.

The main contributions of this work are summarized as follows. (1) We formulate Conversational Stance-Cause Pair Detection (CSCPD) as a new task that extends conversational stance analysis by explicitly modeling the observable conversational evidence for stance expressions. (2) We introduce Cause-CSD, the first high-quality dataset designed for CSCPD, covering both textual and multimodal conversational scenarios. (3) We propose SCD-LM, a unified LLM-based framework that jointly predicts stances and their causes while generating interpretable rationales. (4) Extensive experiments demonstrate that SCD-LM consistently outperforms strong baselines on both subtasks, validating the effectiveness of conversationally grounded stance-evidence modeling.

## 2 Related Work

**Conversational Stance Detection Datasets.** Recent years have seen growing interest in CSD datasets derived from social media comments. Representative resources range from early work such as Lai et al. (2018) to later datasets including SRQ (Villa-Cox et al., 2020), Cantonese-CSD (Li et al., 2023d), the multi-target MT-CSD (Niu

et al., 2024b), the multimodal MmMtCSD (Niu et al., 2024a), and a Chinese conversational CSD dataset (Niu et al., 2025).

While these datasets are valuable for advancing CSD research, they primarily focus on surface-level stance identification. Understanding the causes behind stances, however, is essential for deeper insights into the observable contextual evidence associated with stance expressions, enriching conversational analysis, and improving real-world applications. Unlike emotion-cause analysis, which often relies on explicit expressions (Li et al., 2023b; Cheng et al., 2023; Wang et al., 2023), stance is a cognitive inference derived from reasoning and external knowledge; this implicit nature makes stance-cause detection more challenging.

**Stance Detection Approaches.** Recent advancements in stance detection span traditional machine learning (Mohammad et al., 2016), deep learning (Dey et al., 2018), and pre-trained language models (PLMs) such as BERT (Devlin et al., 2019), which improve performance by modeling nuanced semantics. Fine-tuning PLMs is a strong baseline, while recent methods further enhance adaptability via strategies like stance contrastive learning and target-aware prototypical graph contrastive learning (Liang et al., 2022) or target-based data augmentation for zero-shot stance detection (Li et al., 2023c). More recently, LLMs have further advanced stance detection through enhanced background knowledge integration. LLMs can act as knowledge repositories, accessed via targeted prompting techniques (Zhang et al., 2022, 2023). For example, Cai et al. (2023) introduced a human-in-the-loop system using chain-of-thought prompting, allowing manual refinement of reasoning steps. Another approach combines LLMs with traditional methods: first, an information retrieval system segments large text corpora, and then retrieved knowledge is incorporated into input prompts before feeding into trainable models (Lan et al., 2023; Li et al., 2023a; Ding et al., 2024; Upadhyaya et al., 2025; Dai et al., 2025).

Another line of research emphasizes structure-centric stance modeling. Prior work has shown that reply structure, interaction patterns, and structural embeddings can provide strong signals for stance prediction and stance change modeling (Li et al., 2018; Porco and Goldwasser, 2020; Pick et al., 2022). Compared with these approaches, our task focuses on utterance-level observable contextual evidence within a conversation thread, whereas structure-centric methods capture broader relational grounding for stance.

Another relevant line of research explains stance through interaction structure rather than local semantic evidence alone. These structure-centric methods model stance by leveraging reply graphs, user interaction patterns, and community or alignment signals, showing that relational structure itself can provide strong cues for stance prediction and interpretation. Representative studies, including STEM and prior work by Li, Porco, and Goldwasser as well as Porco and Goldwasser, emphasize that stance may emerge from broader conversational and social structure rather than from a single explicitly stated reason. Compared with these approaches, our task focuses on utterance-level observable contextual evidence within a local conversation thread. Thus, structure-centric modeling and our formulation address different but complementary forms of explanation: the former captures broader relational grounding, whereas CSCPD targets explicit evidence attribution for individual stance expressions.

## 3 Cause-CSD Dataset

**Data Sources.** The Cause-CSD dataset is constructed by adding explicit stance-cause annotations to two recent CSD corpora: MT-CSD (Niu et al., 2024b) and MmMtCSD (Niu et al., 2024a). MT-CSD consists of large-scale, open-domain textual conversations collected from social media platforms, while MmMtCSD augments this setting with multimodal threads that include both textual content and embedded images. Unlike the original datasets, which only provide stance labels, Cause-CSD explicitly labels the observable conversational evidence for stance-related utterances, addressing a key gap in conversational analysis. By leveraging these sources, Cause-CSD supports research on both text-only and multimodal stance–cause detection in conversational environments. Further dataset construction details are provided in Appendix A.

**Dataset Annotation.** Following Li et al. (2023b), we annotate stance-cause relations for all utterances labeled "against" or "favor" in the base datasets, while treating "none" as non-stance and omitting cause annotation for these cases. In total, the MT-CSD dataset contains 8,015 stance-related and 7,861 non-stance utterances, while the MmMtCSD

| Subtask A | | | | | |
|---|---|---|---|---|---|
| | Bitcoin | Tesla | SpaceX | Biden | Trump |
| *Kappa* | 0.69 | 0.62 | 0.73 | 0.83 | 0.70 |
| **Subtask B** | | | | | |
| | Bitcoin | Tesla | Post-T | **Avg.** | |
| *Kappa* | 0.76 | 0.72 | 0.67 | 0.72 | |

Table 1: Annotation Consistency for Cause-CSD. Avg. represents the average consistency.

dataset includes 15,839 stance-related and 5,501 non-stance utterances.

Since most non-stance utterances are unrelated to the stance target, we only annotate stance-labeled utterances for this task. Due to the interactive and non-linear nature of conversations, some stance utterances are linked to multiple cause statements, though most have only one. To accommodate such cases, we set a maximum of five causes per stance utterance. The annotation process follows these main guidelines: (1) The cause for a stance utterance can be the utterance itself, any previous utterance in the conversation, or a combination of both. (2) If the stance is influenced by multiple sources, all relevant utterances are annotated as causes, with a maximum of five causes per stance utterance. (3) If the stance arises from information outside the conversation or references content unrelated to the target, the cause is labeled as none. (4) Only utterances expressing a clear stance towards the target are annotated with causes; "none" instances are ignored for cause annotation. Annotation examples and additional details are provided in Appendix B.

**Annotation Quality Assessment.** Following the annotation guidelines, we established a structured workflow involving twelve natural language processing researchers for data annotation. Before formal annotation, two pilot rounds were conducted to ensure reliability and consistency. Annotations from these rounds were reviewed by two expert annotators to confirm readiness for the task. During the formal annotation process, each instance was annotated by at least two annotators. In cases of disagreement, a second round was conducted to compare differences, and the final cause was determined by an expert annotator. Approximately 5.64% of the records were reviewed by experts to resolve conflicts. Inter-annotator agreement was assessed using Cohen's Kappa (McHugh, 2012), following a binary classification protocol (Wang et al., 2023) to account for the variable number of cause labels per instance. Across all targets and modalities, Kappa scores consistently indicated substantial agreement, with overall average agreement reaching 0.72 (see Table 1 for details).

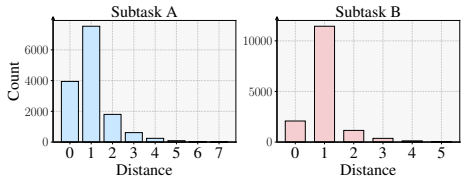


Figure 2: Distances between stance and cause utterances.

**Dataset Analysis.** Cause-CSD[1] contains 34,304 utterances, of which 21,048 are stance-related and 13,256 are labeled as "none" (see Table 2). Each stance-related utterance is paired with one or more cause labels: 13,998 instances have a single cause, while 7,050 are linked to multiple causes (up to five), reflecting the interactive and non-linear nature of real-world conversations. Additionally, 17.19% of stance utterances include the stance utterance itself among their annotated causes, indicating that self-referential stances are common.

We further analyze stance–cause links by measuring the distance between each stance utterance and each of its associated causes (Figure 2). The results show that 19.26% of stance–cause links have distance 0, 64.38% are one utterance apart, 10.01% are two utterances apart, and 6.35% involve longer-range dependencies (three or more utterances apart). This distribution highlights the need to model both local and long-range relationships in conversational stance-cause detection.

Given the limited data for individual stance targets in each subtask, we merge all targets from Subtasks A and B to form a unified dataset. The data is randomly split into training, development, and test sets, following an open-domain experimental protocol (see Table 2 for distribution details). This strategy ensures sufficient diversity and supports robust model evaluation.

## 4 Methodology

### 4.1 Task Definition

Given a conversation $U = \{u_i\}_{i=1}^{n}$, where each $u_i$ may contain text, images, or both, and a target $t$, the task aims to predict a sequence of pairs $Y = \{S_{u_i,t}, C_{u_i,t}\}_{i=1}^{n}$. Here, $S_{u_i,t}$ denotes the

[1]https://github.com/nfq729/Cause-CSD

| Task | | Post | Utterance | None | Same | Number of causes | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | One | Two | Three | Four | Five | Total |
| **Subtask A** | Train | 150 | 11,177 | 5,441 | 2,760 | 2,670 | 1,986 | 776 | 215 | 89 | 5,736 |
| | Dev | 33 | 2,304 | 1,115 | 627 | 527 | 460 | 143 | 42 | 17 | 1,189 |
| | Test | 35 | 2,400 | 1,197 | 559 | 622 | 395 | 130 | 38 | 18 | 1,203 |
| **Subtask B** | Train | 667 | 14,035 | 3,883 | 1,536 | 8,049 | 1,780 | 284 | 34 | 5 | 10,152 |
| | Dev | 144 | 2,307 | 797 | 200 | 1,183 | 269 | 45 | 10 | 3 | 1,510 |
| | Test | 144 | 2,081 | 823 | 214 | 947 | 254 | 46 | 9 | 2 | 1,258 |
| **Total** | | 1,173 | 34,304 | 13,256 | 5,896 | 13,998 | 5,144 | 1,424 | 348 | 134 | 21,048 |
| | | | 100% | 38.64% | 17.19% | 40.81% | 15.00% | 4.15% | 1.01% | 0.39% | 61.36% |

Table 2: Data distribution in the Cause-CSD dataset for Subtasks A and B.

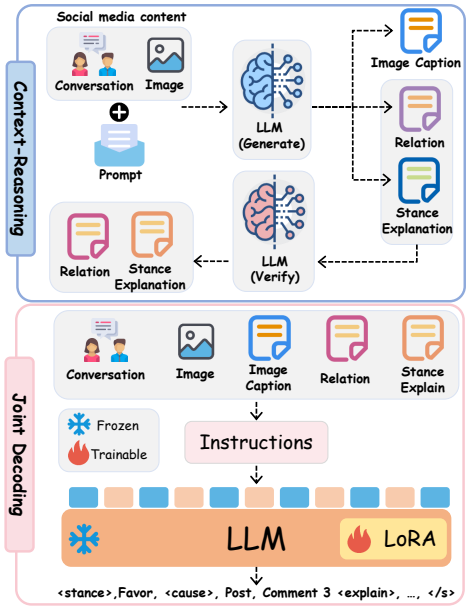


Figure 3: The architecture of our SCD-LM framework.

stance of $u_i$ towards $t$, and $C_{u_i,t}$ identifies the utterance(s), the utterance itself, or multimodal context that serve as the observable contextual evidence for $S_{u_i,t}$ within the thread. We use the term "cause" in this operational sense throughout the paper. If no stance is expressed or no clear evidence is observable in the thread, $C_{u_i,t} = \emptyset$.

## 4.2 Framework Overview

SCD-LM is a two-stage framework (Figure 3) for stance–cause prediction with evidence-aware supervision. Stage 1: Context Reasoning. We use a high-capacity teacher generator to propose stance–cause–rationale tuples and a consistency checker to validate and revise them under the full conversational context, yielding distilled intermediate signals for training. Stage 2: Joint structured decoding. We fine-tune a lightweight decoder on the distilled supervision to jointly predict $\{S_{u_i,t}, C_{u_i,t}\}$ in an autoregressive output, reducing pipeline-style error propagation and encouraging stance–cause consistency. The context-reasoning stage is frozen and does not require task-specific fine-tuning. Only the joint decoding stage is trainable, using the distilled supervision generated by Stage 1.

## 4.3 Context Reasoning

The Context Reasoning stage distills high-quality stance–cause–rationale tuples from complex multi-turn, multimodal conversations to provide explicit supervision for downstream joint decoding. As illustrated in Figure 3, this stage consists of four steps: Image Caption, Relation, Stance Explanation, and Verify. These intermediate signals (i) convert visual evidence into a text proxy and (ii) make the stance–cause link explicit and verifiable under the full context. For each utterance $u_i$ and target $t$, we conduct the following procedure:

**Image Caption.** For conversations containing images, we generate a descriptive caption for each image. The LLM produces a target-aware caption $c_i = M_{\text{LLM},1}(U, t)$ based on the conversation and image, focusing on content relevant to $t$. This caption serves as a textual proxy for visual evidence, enabling subsequent relation extraction, explanation, and joint decoding to operate on a unified text context. The prompt is as follows:

> **Input Data:** $U, t$
> **Prompt:** Generate one sentence describing the parts of the image most relevant to the target and the conversation.
> **Expected Output:** Image caption.

**Relation.** For the candidate utterance $u_i$, the model analyzes its semantic relationships with the target entity, preceding utterances, and associated images. Instead of directly selecting a cause, this step prompts the LLM to determine the type and direction of relations between $u_i$ and other context elements–such as agreement, contrast, support, query, or reference–thereby constructing a relational graph for the conversation. Specifically, we obtain all identified semantic relations as $u_i^r = M_{\text{LLM},2}(U, t, u_i)$, which provide rich struc-

tural cues for subsequent stance-cause pairing and explanation. The prompt is as follows:

**Input Data:** $U, t, u_i$
**Prompt:** Given the conversation, the target entity, and the current comment, identify and describe the semantic relations between this comment and (1) the target entity, (2) each previous utterance, and (3) image content. For each relation, specify the type (e.g., support, contrast, query, reference, none) and direction.
**Expected Output:** A set of relation types (with direction) between $u_i$ and each context element.

**Stance Explanation.** Given the target $t$, the candidate utterance $u_i$, the identified cause $C_{u_i,t}$, and the stance $S_{u_i,t}$, we prompt the LLM to generate a concise, natural-language rationale $u_i^e$ that explains why the identified cause is annotated as the observable contextual evidence for the stance in $u_i$. During training, the annotated cause is used to supervise rationale generation, enabling the model to learn from gold-standard explanations. During inference, the model relies on its own predicted cause to generate explanations, ensuring a fair evaluation of end-to-end performance. Specifically, the rationale is generated as $u_i^e = M_{\text{LLM},3}(U, t, u_i, C_{u_i,t}, S_{u_i,t})$. The prompt is as follows:

**Input Data:** $U, t, C_{u_i,t}, S_{u_i,t}$
**Prompt:** Given the target entity, the candidate utterance (with predicted stance), and the identified cause, provide a brief explanation (no more than 60 words) for why the identified cause is the observable contextual evidence associated with the stance.
**Expected Output:** A concise natural-language rationale explaining the evidence-stance association in $u_i$.

**Verify.** To ensure logical consistency and correctness, we prompt the LLM to verify the relation–rationale pair $(u_i^r, u_i^e)$ within the full conversational context and target. Specifically, the LLM receives the entire conversation $U$, target entity $t$, the set of identified relations $u_i^r$, and the generated rationale $u_i^e$, and is required to check whether $u_i^e$ is coherent and factually correct with respect to $u_i^r$ and the overall context. If any inconsistency or error is found, the LLM proposes a revision; otherwise, it confirms the pair as valid. Formally, this verification is performed as $(u_i^r, u_i^e) = M_{\text{LLM},4}(U, t, u_i^r, u_i^e)$. The prompt is as follows:

**Input Data:** $U, t, u_i^r, u_i^e$
**Prompt:** Given the conversation, target entity, the identified relations, and the rationale, verify whether the rationale is logically consistent and factually correct with respect to the relations and the context. If consistent, reply VALID; otherwise, return REVISE and the corrected pair.
**Expected Output:** Either "VALID" or a revised $(u_i^r, u_i^e)$ pair.

### 4.4 Joint Decoding

In the second stage, we employ a joint decoding strategy to simultaneously predict the stance and its corresponding cause for each utterance, leveraging the rationale-augmented data generated from the context-reasoning stage. This unified approach mitigates error propagation and enforces logical consistency between stance and cause. As illustrated in Figure 3, we first construct an instruction set that organizes the conversation, associated images, generated image captions, semantic relations, and stance explanations into a unified, structured input format. These instructions provide the model with rich, context-aware cues for joint stance-cause reasoning. The instruction-formatted data is then used to fine-tune a parameter-efficient decoder (LLM with LoRA (Hu et al., 2022)). For the candidate utterance $u_i$, given the full conversation $U$, the target entity $t$, and (where applicable) image captions, the model generates both the stance $S_{u_i,t}$ and the cause $C_{u_i,t}$ in an autoregressive output sequence. Optionally, the explanation $u_i^e$ can also be generated for interpretability. Formally, the joint prediction is defined as:

$$(S_{u_i,t}, C_{u_i,t}, u_i^e) = M_{\text{Dec}}(U, t, u_i)$$

where $M_{\text{Dec}}$ is the parameter-efficient decoder trained on rationale-augmented supervision.

During training, we optimize the conditional likelihood of the gold stance, cause, and explanation given the input context:

$$\mathcal{L} = -\sum_{i=1}^{n} \log P(S_{u_i,t}, C_{u_i,t}, u_i^e \mid U, t, u_i)$$

At inference time, the model receives a new conversation and target, and jointly decodes the stance–cause pair $(S_{u_i,t}, C_{u_i,t})$ (and optionally the rationale $u_i^e$) for each utterance.

During inference, only the predicted stance $S_{u_i,t}$ and cause $C_{u_i,t}$ are evaluated. The explanation $u_i^e$ is excluded to ensure true end-to-end performance.

By generating stance and cause in a unified sequence, the model captures dependencies and mutual constraints between them, resulting in more robust and logically consistent predictions. This joint decoding framework offers several advantages: (1) it avoids error accumulation common in pipeline approaches; (2) it ensures stance and cause predictions are mutually coherent; (3) it enables the

model to leverage explicit reasoning traces for improved generalization, especially in long-range or multimodal cases.

## 5 Experiments

### 5.1 Experimental Settings

**Baseline Methods.** To benchmark our method on both subtask A (text-only) and subtask B (multimodal), we compare against strong baselines, including neural networks, PLMs, and LLMs. **Subtask A (Text-only):** LSTM (Hochreiter, 1997), Joint-GCN/Xatt (Li et al., 2023b), BERT (Devlin et al., 2019), CD-MRC (Cheng et al., 2023), Llama-70b, ChatGPT (GPT-3.5/4/4o), Claude 3.5. **Subtask B (Multimodal):** BERT+ViT (Liang et al., 2024), MECPE-2steps (Wang et al., 2023), MER-MCE (Cheng et al., 2024), MLLM-SD (Niu et al., 2024a), ChatGPT (GPT-4 Vision, 4o), Claude 3.5. LLMs are evaluated using prior prompting protocols (Lei et al., 2024; Cheng et al., 2024).

**Evaluation Metrics.** Following prior work (Li et al., 2021; Wang et al., 2023), we use the average F1 of informative classes (Stance-F1) for stance evaluation. For cause identification, we report the F1 scores for both cause extraction (Cause-F1) and joint stance-cause pair extraction (Pair-F1). Detailed definitions are in Appendix C.

**Implementation Details.** For the context-reasoning stage, we utilize LLMs as external teachers: GPT-4o is employed for semantic relation and rationale generation, while GPT-4.1 is used for verification. For the joint decoding stage, we use different models for each subtask. For Subtask A (text-only), we fine-tune Llama 3.1-8B; for Subtask B (multimodal), we fine-tune Llama 3.2-Vision-11B. Both models are adapted using LoRA (Hu et al., 2022) (rank 8, $\alpha$=16). Fine-tuning is implemented with the Llama-Factory framework (Zheng et al., 2024) built on HuggingFace Transformers. All experiments are conducted on eight NVIDIA A100 GPUs (40GB each), and results are averaged over three random seeds for robustness.

### 5.2 Experimental Results

**Open-domain Experimental Results.** In Table 3, we compare the performance of our proposed SCD-LM with a broad range of baselines on the Cause-CSD dataset. The results demonstrate that SCD-LM achieves state-of-the-art performance on both subtasks. For text-only conversations (Subtask A), SCD-LM delivers a Stance-F1 of 59.61 and a Cause-F1 of 54.96, outperforming classic models (LSTM, Joint-GCN) and even powerful LLM baselines. Notably, SCD-LM's stance-cause Pair-F1 reaches 34.57, which is about 3.56 points higher than the best baseline (GPT-4o, 31.01). This indicates SCD-LM's superior ability to jointly identify stances and their causes compared to pipeline or single-stage methods. In multimodal conversations (Subtask B), SCD-LM similarly leads in all metrics. It attains 73.10 Stance-F1 and 55.54 Cause-F1, surpassing strong multimodal models like MER-MCE and even GPT-4 Vision. The Pair-F1 of SCD-LM is 42.78, marking a new high in the accuracy of stance-cause pairing for multimodal data. These improvements confirm that our two-stage framework, which integrates LLM-driven reasoning with joint decoding, effectively leverages both textual and visual context. Llama-2-70B is not uniformly weak; rather, its relatively competitive Cause-F1 but lower Stance-F1 and Pair-F1 suggest that prompting-based LLMs struggle more with strict evidence localization and index-based pairing than with free-form reasoning.

**Impact of Cause Distance.** Table 4 further analyzes performance by the distance between a stance utterance and its cause. These results indicate a clear trend that all models find it increasingly difficult to detect causes as the distance grows, but SCD-LM's advantage becomes more pronounced on longer-range dependencies. For Subtask A, when the cause is in the same utterance (distance = 0), SCD-LM achieves Stance-F1 61.44 and Pair-F1 37.65, outperforming the best baseline (GPT-4o) by approximately 4 points in Pair-F1. As the gap extends to intervening utterances (distance = 1), the performance of all methods drops; however, SCD-LM degrades more gracefully. At distance $\geq$ 2, SCD-LM still attains a Pair-F1 of 33.21, whereas the strongest baseline remains below 29. This margin of over 4 points at longer distances highlights SCD-LM's superior ability to retain and reason over extended conversational contexts. A similar pattern holds in Subtask B. SCD-LM consistently ranks first or tied for first in Stance-F1 and Cause-F1 across all distance segments. Notably, for distant causes in multimodal threads, SCD-LM achieves Pair-F1 41.26 (at distance $\geq$ 2), about 2.3 points higher than the closest competitor. We also observe that while certain baselines perform

| Task | METHOD | Against-F1 | Favor-F1 | Stance-F1 | Cause-P | Cause-R | Cause-F1 | Pair-P | Pair-R | Pair-F1 |
|---|---|---|---|---|---|---|---|---|---|---|
| Subtask A | LSTM | 40.25 | 52.67 | 46.46 | 34.97 | 38.54 | 36.67 | 14.12 | 19.69 | 16.45 |
| | Joint-GCN | 44.08 | 52.71 | 48.40 | 35.19 | 39.98 | 37.43 | 17.32 | 20.92 | 18.95 |
| | Joint-Xatt | 47.72 | 53.40 | 50.56 | 35.94 | 37.81 | 36.85 | 15.99 | 21.05 | 18.17 |
| | Bert | 56.92 | 55.73 | 56.33 | 42.45 | 49.85 | 45.85 | 19.87 | 26.79 | 22.82 |
| | CD-MRC | 58.39 | 57.45 | 57.92 | 41.45 | 48.96 | 44.89 | 20.95 | 25.82 | 23.13 |
| | Llama 2-70b | 48.16 | 48.19 | 48.18 | 44.95 | 52.13 | 48.27 | 22.31 | 27.84 | 24.77 |
| | GPT-3.5 | 49.09 | 45.51 | 47.30 | 46.51 | 53.14 | 49.60 | 26.61 | 28.00 | 27.29 |
| | GPT-4 | 51.26 | 50.92 | 51.09 | 46.93 | 55.18 | 50.72 | 27.83 | 29.33 | 28.56 |
| | GPT-4o | 53.48 | 55.90 | 54.69 | 48.72 | 55.89 | 52.06 | 29.85 | 32.27 | 31.01 |
| | Claude 3.5 | 54.63 | 56.10 | 55.37 | 47.35 | 55.03 | 50.90 | 29.10 | 31.64 | 30.32 |
| | SCD-LM | **59.02** | **60.20** | **59.61** | **52.03** | **58.25** | **54.96** | **33.21** | **36.05** | **34.57** |
| Subtask B | BERT+ViT | 63.48 | 68.87 | 66.18 | 35.09 | 43.34 | 38.78 | 28.03 | 33.60 | 30.56 |
| | MECPE-2steps | 60.03 | 65.12 | 62.58 | 42.60 | 47.06 | 44.72 | 32.61 | 37.81 | 35.02 |
| | MER-MCE | 66.35 | 68.06 | 67.21 | 48.21 | 50.01 | 49.09 | 34.45 | 47.38 | 39.89 |
| | MLLM-SD | 68.47 | 71.32 | 69.90 | 50.47 | 55.51 | 52.87 | 34.67 | 48.55 | 40.45 |
| | GPT4-Vision | 68.15 | 72.45 | 70.30 | 49.93 | 55.63 | 52.63 | 33.83 | 48.73 | 39.94 |
| | GPT-4o | 69.20 | 73.05 | 71.13 | 49.64 | 56.70 | 52.94 | 34.84 | 49.10 | 40.76 |
| | Claude 3.5 | 68.70 | 72.90 | 70.80 | 50.12 | 56.05 | 52.92 | 34.60 | 48.40 | 40.35 |
| | SCD-LM | **71.30** | **74.90** | **73.10** | **52.87** | **58.49** | **55.54** | **37.58** | **49.66** | **42.78** |

Table 3: Open-domain experimental results on the Cause-CSD dataset.

| Task | METHOD | Distance=0 | | | Distance=1 | | | Distance≥ 2 | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | Stance-F1 | Cause-F1 | Pair-F1 | Stance-F1 | Cause-F1 | Pair-F1 | Stance-F1 | Cause-F1 | Pair-F1 |
| Subtask A | LSTM | 48.67 | 39.23 | 18.89 | 46.23 | 36.56 | 16.23 | 45.89 | 30.59 | 12.97 |
| | Joint-GCN | 49.83 | 40.64 | 20.73 | 50.31 | 36.74 | 17.21 | 46.74 | 33.27 | 13.19 |
| | Joint-Xatt | 51.81 | 41.56 | 21.56 | 49.42 | 35.67 | 17.65 | 48.01 | 32.22 | 15.89 |
| | Bert | 55.47 | 47.93 | 25.32 | 57.67 | 45.11 | 21.46 | 55.23 | 41.54 | 18.36 |
| | CD-MRC | 58.25 | 47.26 | 24.79 | 55.64 | 44.04 | 22.23 | 54.04 | 42.74 | 19.88 |
| | Llama 2-70b | 50.80 | 50.74 | 29.91 | 48.04 | 47.69 | 23.55 | 47.90 | 45.78 | 20.46 |
| | GPT-3.5 | 46.03 | 52.60 | 30.19 | 49.82 | 48.21 | 27.61 | 44.98 | 44.92 | 23.65 |
| | GPT-4 | 54.18 | 51.81 | 32.90 | 50.72 | 50.10 | 30.24 | 48.82 | 45.71 | 25.97 |
| | GPT-4o | 57.62 | 53.38 | 33.48 | 54.13 | 50.20 | 31.44 | 52.42 | 46.23 | 28.37 |
| | Claude 3.5 | 57.19 | 52.16 | 32.79 | 53.70 | 49.76 | 29.66 | 51.08 | 44.90 | 27.44 |
| | **SCD-LM** | **61.44** | **58.82** | **37.65** | **60.55** | **57.29** | **35.10** | **56.01** | **52.37** | **33.21** |
| Subtask B | BERT+ViT | 65.27 | 41.60 | 32.11 | 68.03 | 38.02 | 30.39 | 63.68 | 35.95 | 28.71 |
| | MECPE-2steps | 65.45 | 47.69 | 39.18 | 61.85 | 42.24 | 35.11 | 62.66 | 40.03 | 32.20 |
| | MER-MCE | 67.34 | 50.98 | 41.04 | 66.97 | 48.91 | 39.86 | 66.21 | 44.24 | 35.21 |
| | MLLM-SD | 71.76 | 58.06 | 42.73 | 69.05 | 52.77 | 39.32 | 67.72 | 48.55 | 37.21 |
| | GPT4-Vision | 72.03 | 55.88 | 40.81 | 68.56 | 52.26 | 37.24 | 69.58 | 46.96 | 38.07 |
| | GPT-4o | 72.90 | 57.03 | 42.27 | 69.34 | 54.58 | 41.78 | 68.05 | 49.19 | 38.96 |
| | Claude 3.5 | 71.20 | 56.18 | 41.10 | 68.90 | 54.04 | 38.41 | 67.51 | 47.35 | 36.32 |
| | **SCD-LM** | **74.92** | **60.42** | **44.13** | **70.15** | **55.73** | **44.51** | **70.60** | **50.02** | **41.26** |

Table 4: The performance evaluation presents a comparison of the scores for different models in instances where the distance is 0, 1, and ≥ 2.

strongly on immediate, image-grounded cases (distance = 0), their performance drops sharply with increased distance. In contrast, SCD-LM remains robust even when causal utterances are distant or appear in images, highlighting the benefits of our rationale-augmented joint modeling.

**Ablation Study.** We conducted an ablation study to systematically examine the contributions of individual components within our SCD-LM framework. Figure 4 clearly demonstrates that removing the Context Reasoning module significantly impairs performance, suggesting its vital role in capturing essential context for stance-cause detection in both textual (Subtask A) and multimodal (Subtask B) settings. Additionally, omitting the Stance Explanation component also leads to notable reductions, highlighting the importance of explicit rationales for linking stances and causes coherently. Removing either the Relation extraction or Verification

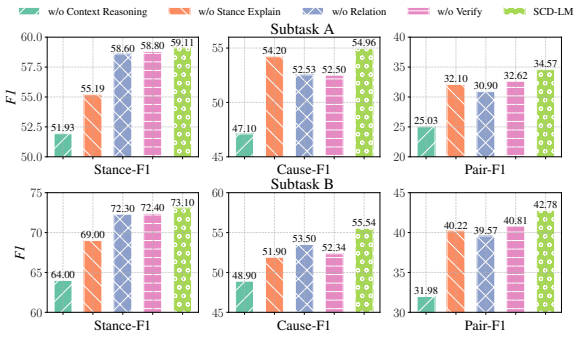


Figure 4: Ablation study results on SCD-LM.

module similarly decreases performance, though these impacts are relatively moderate compared to the aforementioned components. These findings collectively indicate that each component in our framework is integral, with Context Reasoning and Stance Explanation being particularly influential in ensuring accurate, logically consistent predictions.

### 5.3 Analysis and Discussion

**Q1: Is It Necessary to Construct Stance-Cause Data?** Constructing specialized stance-cause annotated data is essential, as it enables models to capture the contextual grounding behind stance expressions, moving beyond surface-level stance classification alone. Traditional stance detection datasets typically focus only on stance polarity, while overlooking the observable conversational evidence that supports or contextualizes a stance in multi-turn interactions. Our Cause-CSD dataset fills this critical gap by explicitly annotating stance-cause relationships in this operational sense, facilitating more informative analyses and promoting the development of models capable of capturing richer semantic interactions within conversations.

**Q2: Is the SCD-LM Mechanism Reasonable?** The proposed two-stage SCD-LM architecture demonstrates substantial advantages in addressing the complexities of joint stance-cause detection. By first leveraging large language models to extract structured rationales (Context Reasoning) and subsequently jointly decoding stance-cause pairs, SCD-LM effectively mitigates common issues such as error propagation observed in pipeline methods. In Figure 4, the ablation studies further validate the validity of our design choices, showing consistent performance degradation when essential modules (especially Context Reasoning and Stance Explanation) are removed. These findings confirm that the structured reasoning combined with unified decoding is both practical and effective.

**Q3: What Are SCD-LM's Current Limitations?** Despite its strong performance, SCD-LM exhibits several limitations that warrant further exploration. First, while Context Reasoning significantly enhances performance, it relies heavily on external large language models, making the system resource-intensive. Additionally, the performance in long-range cause detection scenarios, though superior to baselines, still exhibits considerable room for improvement. Lastly, handling multiple causes simultaneously remains challenging, as our current model primarily emphasizes single-cause extraction in practice. Future research could investigate strategies to better integrate multiple-cause predictions and improve efficiency and scalability for broader application scenarios.

## 6 Conclusion

We introduced CSCPD, a new task that jointly predicts stance polarity and its observable contextual evidence in multi-turn conversations. To support this task, we built Cause-CSD, the first large-scale dataset with explicit stance–cause annotations, and proposed SCD-LM, a unified framework that combines context reasoning with joint decoding to improve stance–cause consistency. Experiments on both text-only and multimodal settings show that SCD-LM consistently outperforms strong baselines, especially for long-range and image-grounded cause detection. Future directions include improving multi-cause extraction and reducing reliance on external LLMs.

## Acknowledgments

This work has been supported by the National Key R&D Program of China (No. U25B2042), the Guangdong S&T Program (2025B0101130002), the National Natural Science Foundation of China (No. 62306184), the Natural Science Foundation for Top Talents of SZTU (Nos. GDRC202518 and GDRC202320), the Shenzhen Science and Technology Program (Nos. RCBS20231211090548077 and JCYJ20240813113218025), the Guangdong Basic and Applied Basic Research Foundation (2026A1515010133), and the Project for Improving Scientific Research Capabilities of Key Construction Disciplines in Guangdong Province (No. 2025ZDJS039).

## Limitations

Our work has several limitations. First, the term "cause" is used here in an operational sense to denote observable contextual evidence for a stance expression within the conversation thread, rather than the speaker's latent beliefs, identity alignment, affective motivations, or broader belief formation process. Accordingly, strong performance on this benchmark should be interpreted as recovering annotator-consistent evidence in discourse, not as uncovering the true causal mechanism of stance formation. This setting may still involve post hoc rationalization risk, especially when stance expressions are habitual, affective, socially aligned, or only weakly grounded in the observable conversation. In addition, while Cause-CSD annotates up to five causes per stance utterance and our framework supports multi-cause supervision, accurately recovering all relevant causes remains challenging,

especially when evidence is implicit, distributed across long conversational contexts, or partially grounded in images. Finally, the context-reasoning stage relies on external LLMs for relation and rationale generation as well as verification, which increases computational cost and may affect reproducibility if model versions or access conditions change.

## Ethical Considerations

Our dataset is constructed by extending two existing resources, MT-CSD (Niu et al., 2024b) and MmMtCSD (Niu et al., 2024a), which were originally collected from public Reddit threads via the official Reddit API and have been used in prior research. We use these datasets in accordance with their original usage conditions and the platform's policies, and we will ensure that any release of our derived annotations follows the same redistribution constraints.

The additional annotations in this work were produced by trained annotators with relevant background knowledge. In total, twelve annotators participated in the annotation process. Each annotator is paid $6.5 per hour (above the average local payment of similar jobs). The entire annotation process lasted 5 months, and the average annotation time of the twelve annotators was 510 hours. All annotators were informed of the research purpose and participated voluntarily.

For the LLM-based components (used only as external teachers for relation and rationale generation and verification), we access models through official interfaces and strictly follow the providers' usage policies. We do not intentionally submit private or sensitive user information, and model-generated rationales are treated as research artifacts rather than verified facts.

Finally, we note that stance and cause inference technologies may be misused in sensitive applications such as targeted persuasion or large-scale monitoring. We therefore position this work strictly for research purposes and emphasize the importance of human oversight and responsible use in any downstream deployment.

## References


Isabelle Augenstein, Tim Rocktäschel, Andreas Vlachos, and Kalina Bontcheva. 2016. Stance detection with bidirectional conditional encoding. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 876–885, Austin, Texas. Association for Computational Linguistics.

Zefan Cai, Baobao Chang, and Wenjuan Han. 2023. Human-in-the-loop through chain-of-thought. *arXiv preprint arXiv:2306.07932*.

Zebang Cheng, Fuqiang Niu, Yuxiang Lin, Zhi-qi Cheng, Xiaojiang Peng, and Bowen Zhang. 2024. 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, Mexico City, Mexico. Association for Computational Linguistics.

Zifeng Cheng, Zhiwei Jiang, Yafeng Yin, Cong Wang, Shiping Ge, and Qing Gu. 2023. A consistent dual-mrc framework for emotion-cause pair extraction. *ACM Trans. Inf. Syst.*, 41(4).

Genan Dai, Jiayu Liao, Sicheng Zhao, Xianghua Fu, Xiaojiang Peng, Hu Huang, and Bowen Zhang. 2025. Large language model enhanced logic tensor network for stance detection. *Neural Networks*, 183:106956.

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.

Kuntal Dey, Ritvik Shrivastava, and Saroj Kaushik. 2018. Topical stance detection for twitter: A two-phase lstm model using attention. In *European Conference on Information Retrieval*, pages 529–536. Springer.

Daijun Ding, Rong Chen, Liwen Jing, Bowen Zhang, Xu Huang, Li Dong, Xiaowen Zhao, and Ge Song. 2024. Cross-target stance detection by exploiting target analytical perspectives. *arXiv preprint arXiv:2401.01761*.

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.

S Hochreiter. 1997. Long short-term memory. *Neural Computation MIT-Press*.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Dilek Küçük and Fazli Can. 2020. Stance detection: A survey. *ACM Computing Surveys (CSUR)*, 53(1):1–37.

Mirko Lai, Viviana Patti, Giancarlo Ruffo, and Paolo Rosso. 2018. Stance evolution and twitter interactions in an italian political debate. In *Natural Language Processing and Information Systems - 23rd International Conference on Applications of Natural Language to Information Systems, NLDB 2018, Paris, France, June 13-15, 2018, Proceedings*, volume 10859 of *Lecture Notes in Computer Science*, pages 15–27. Springer.

Xiaochong Lan, Chen Gao, Depeng Jin, and Yong Li. 2023. Stance detection with collaborative role-infused llm-based agents. *arXiv preprint arXiv:2310.10467*.

Shanglin Lei, Guanting Dong, Xiaoping Wang, Keheng Wang, and Sirui Wang. 2024. Instructerc: Reforming emotion recognition in conversation with a retrieval multi-task llms framework. *Preprint*, arXiv:2309.11911.

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, Xingwei Liang, Kam-Fai Wong, and Ruifeng Xu. 2025. Mitigating biases of large language models in stance detection with counterfactual augmented calibration. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7075–7092.

Chang Li, Aldo Porco, and Dan Goldwasser. 2018. Structured representation learning for online debate stance prediction. In *Proceedings of the 27th international conference on computational linguistics*, pages 3728–3739.

Wei Li, Yang Li, Vlad Pandelea, Mengshi Ge, Luyao Zhu, and Erik Cambria. 2023b. Ecpec: Emotion-cause pair extraction in conversations. *IEEE Transactions on Affective Computing*, 14(3):1754–1765.

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.

Yingjie Li, Chenye Zhao, and Cornelia Caragea. 2023c. Tts: A target-based teacher-student framework for zero-shot stance detection. In *Proceedings of the ACM Web Conference 2023*, pages 1500–1509.

Yupeng Li, Haorui He, Shaonan Wang, Francis CM Lau, and Yunya Song. 2023d. Improved target-specific stance detection on social media platforms by delving into conversation threads. *IEEE Transactions on Computational Social Systems*.

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. *arXiv preprint arXiv:2402.14298*.

Bin Liang, Qinlin Zhu, Xiang Li, Min Yang, Lin Gui, Yulan He, and Ruifeng Xu. 2022. Jointcl: a joint contrastive learning framework for zero-shot stance detection. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 81–91. Association for Computational Linguistics.

Mary L McHugh. 2012. Interrater reliability: the kappa statistic. *Biochemia medica*, 22(3):276–282.

Saif Mohammad, Svetlana Kiritchenko, Parinaz Sobhani, Xiao-Dan Zhu, and Colin Cherry. 2016. Semeval-2016 task 6: Detecting stance in tweets. In *Proceedings of the 10th International Workshop on Semantic Evaluation, SemEval@NAACL-HLT, San Diego, CA, USA, June 16-17*, pages 31–41.

Saif M. Mohammad, Parinaz Sobhani, and Svetlana Kiritchenko. 2017. Stance and sentiment in tweets. *ACM Trans. Internet Technol.*, 17(3).

Fuqiang Niu, Zebang Cheng, Xianghua Fu, Xiaojiang Peng, Genan Dai, Yin Chen, Hu Huang, and Bowen Zhang. 2024a. Multimodal multi-turn conversation stance detection: A challenge dataset and effective model. In *ACM Multimedia 2024*.

Fuqiang Niu, Min Yang, Ang Li, Baoquan Zhang, Xiaojiang Peng, and Bowen Zhang. 2024b. 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*, WWW '25, page 769–772, New York, NY, USA. Association for Computing Machinery.

Ron Korenblum Pick, Vladyslav Kozhukhov, Dan Vilenchik, and Oren Tsur. 2022. Stem: unsupervised structural embedding for stance detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 11174–11182.

Aldo Porco and Dan Goldwasser. 2020. Predicting stance change using modular architectures. In *Proceedings of the 28th international conference on computational linguistics*, pages 396–406.

Rudra Ranajee Saha, Laks VS Lakshmanan, and Raymond T Ng. 2024. Stance detection with explanations. *Computational Linguistics*, 50(1):193–235.

Swapna Somasundaran and Janyce Wiebe. 2010. Recognizing stances in ideological on-line debates. In *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, pages 116–124, Los Angeles, CA. Association for Computational Linguistics.

Apoorva Upadhyaya, Wolfgang Nejdl, and Marco Fisichella. 2025. Interpretable zero-shot stance detection with proactive content intervention. *Information Processing Management*, 62(6):104223.

Ramon Villa-Cox, Sumeet Kumar, Matthew Babcock, and Kathleen M Carley. 2020. Stance in replies and quotes (srq): A new dataset for learning stance in twitter conversations. *arXiv preprint arXiv:2006.00691*.

Fanfan Wang, Zixiang Ding, Rui Xia, Zhaoyu Li, and Jianfei Yu. 2023. Multimodal emotion-cause pair extraction in conversations. *IEEE Transactions on Affective Computing*, 14(3):1832–1844.

Bowen Zhang, Daijun Ding, and Liwen Jing. 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, Yangyang Li, and Liwen Jing. 2023. Investigating chain-of-thought with chatgpt for stance detection on social media. *arXiv preprint arXiv:2304.03087*.

Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, and Zheyan Luo. 2024. LlamaFactory: Unified efficient fine-tuning of 100+ language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 400–410, Bangkok, Thailand. Association for Computational Linguistics.

## A Datasets Detail

We construct Cause-CSD by adding explicit cause annotations to two recent conversational stance detection corpora. The first is the MT-CSD (Niu et al., 2024b) text-only dataset, containing 15,876 English posts and comments from Reddit spanning five controversial targets: *Tesla*, *SpaceX*, *Donald Trump*, *Joe Biden*, and *Bitcoin*. These data were originally collected via the official Reddit API, ensuring authentic, multi-turn discussions with substantial context. The second source is the MmMtCSD (Niu et al., 2024a) multimodal dataset, which consists of 21,340 Reddit posts and comments paired with images, focusing on three target settings: two specific targets (*Tesla* and *Bitcoin*) and a more general Post-Target (*Post-T*) setting where each post serves as its own target, yielding a diverse range of topics. A detailed breakdown of both datasets is provided in Table 5.

We extend the full content of these datasets with cause annotations, resulting in Cause-CSD, which covers all multi-turn conversation threads from the original corpora while preserving their stance labels and conversational structure.

| Dataset | Target | Samples and Proportion of Labels | | | |
|---|---|---|---|---|---|
| | | Against | Favor | None | Total |
| MT-CSD | **Bitcoin** | 1,324 | 869 | 1,192 | 3,385 |
| | **Tesla** | 1,146 | 477 | 2,068 | 3,691 |
| | **SpaceX** | 298 | 595 | 1,162 | 2,055 |
| | **Biden** | 352 | 1,186 | 1,409 | 2,947 |
| | **Trump** | 1,667 | 207 | 1,924 | 3,798 |
| MmMtCSD | **Tesla** | 2,211 | 2,531 | 1,558 | 6,300 |
| | **Bitcoin** | 1,284 | 4,550 | 2,314 | 8,148 |
| | **Post-T** | 2,008 | 3,255 | 1,629 | 6,892 |
| **Total** | | 10,290 | 13,670 | 13,256 | 37,216 |

Table 5: Data distribution of MT-CSD dataset and MmMtCSD dataset.

## B Annotation Detail

Following Li et al. (2023b), we annotate stance-cause relations for all utterances labeled "against" or "favor" in the base datasets, while treating "none" as non-stance and omitting cause annotation for these cases. In total, the MT-CSD dataset contains 8,015 stance-related and 7,861 non-stance utterances, while the MmMtCSD dataset includes 15,839 stance-related and 5,501 non-stance utterances.

Since most non-stance utterances are unrelated to the stance target, we only annotate stance-labeled utterances for this task. Due to the interactive and non-linear nature of conversations, some stance utterances are linked to multiple cause statements, though most have only one. To accommodate such cases, we set a maximum of five causes per stance utterance. The annotation process follows these main guidelines:

(1) Cause within the same or previous utterances. The cause for a stance utterance can be the utterance itself, any previous utterance in the conversation, or a combination of both. For example, in Table 6, the stance of Comment 1 is labeled as Against, and its cause includes both the Post and Comment 1 itself. The comment expresses dissatisfaction with Britain's drug policy while also referring to the Post about Biden's marijuana pardon, making both utterances part of the cause.

(2) Multiple causes. If the stance is influenced by multiple sources, all relevant utterances are annotated as causes, with a maximum of five causes

| Target | Example |
|---|---|
| *Joe Biden* | *Post*: Biden to pardon all prior federal offenses of simple marijuana possession. [Stance: *None*] *Comment 1*: Meanwhile in Britain, it was reported this week that Tory Police Commissioners want to make cannabis a Class A drug, bringing it in line with cocaine and heroin. What a shitpit we have become. [Stance: *Against*, Cause: *Post, Comment 1*] |

Table 6: Dataset annotation example.

per stance utterance. As illustrated in Table 6, Comment 1 draws its stance from two distinct utterances (the Post and itself), showing that multiple pieces of context can jointly serve as evidence for a stance expression.

(3) No identifiable cause (None). The cause label is set to None when the stance is inferred from the utterance, but the relevant evidence is not observable within the conversation or references content unrelated to the stance target. For example, if the stance target is Joe Biden and the utterance is "A dark Brandon appears", the stance is Against, but no specific cause can be found in the conversation context. In this case, the cause is labeled as None, since the remark is nonsensical and lacks a relevant explanation within the thread.

(4) Non-stance utterances. Only utterances expressing a clear stance towards the target (i.e., labeled as "against" or "favor") are annotated with causes; utterances labeled as "none" are ignored for cause annotation.

## C Evaluation metrics

For the evaluation of stance, we adopt the Stance-F1 metric, consistent with the approaches in Li et al. (2021) and Mohammad et al. (2017). Stance-F1 represents the average F1 score computed for the "against" and "favor" stances, denoted as against-F1 and favor-F1, respectively.

For the evaluation of cause, we conducted separate assessments for both cause and stance-cause extraction. Consistent with Wang et al. (2023), we used Precision, Recall, and F1 scores as evaluation metrics. Specifically, Cause-P, Cause-R, and Cause-F1 were used to evaluate the performance of cause extraction, while Pair-P, Pair-R, and Pair-F1 were employed to assess the effectiveness of stance-cause extraction. The calculations for Pair-P, Pair-R, and Pair-F1 are as follows:

$$P = \frac{\sum correct\_pairs}{\sum predicted\_pairs} \quad (1)$$

$$R = \frac{\sum correct\_pairs}{\sum annotated\_pairs} \quad (2)$$

$$F_1 = \frac{2 \times P \times R}{P + R} \quad (3)$$

where *predicted_pairs* denotes the number of stance-cause pairs predicted by the model, *annotated_pairs* refers to the total number of stance-cause pairs annotated in the dataset, and *correct_pairs* indicates the number of pairs that are both annotated and correctly predicted as stance-cause pairs.