Disagreement-Aware Robust Training for Multi-modal Stance Detection under Model-internal Decision Inconsistency

Anonymous ACL submission

Abstract

Stance detection is usually studied in a single modality, focusing on semantic understanding. On social media, however, stance is expressed in more diverse ways. Multi-modal stance detection (MSD) leverages paired text and images to enrich stance expressions but also introduces a challenge in fusing different modalities. Interestingly, our study uncovers an instructive model-internal state: the text and image encoders can yield inconsistent stance decisions, even when the input pair conveys a unified stance. We term this measurable state as *Modal Decision Disagreement* (MDD). Under this, standard training only supervises the final fused output, it does not constrain how the model should handle these conflicting internal signals. Thus, simple averaging or alignment-oriented fusion often turns into a wrong or compromise prediction. To address this, we propose **DART**, a disagreement-aware robust training framework. Specifically, we utilize a decision-level auxiliary head to regularize the fused predictor against branch disagreement. Moreover, to further improve robustness to such inconsistencies, we apply a text stance-flip perturbation that creates deliberately conflicting training instances. Together, they make fusion more stable under branch-level disagreement. Across all five MSD benchmarks, we improve both in-target and zero-shot performance, with the largest gains when the model exhibits MDD.

1 Introduction

Stance detection has traditionally been studied through the lens of uni-modal semantic analysis, focused on inferring attitudes primarily from textual content (Küçük and Can, 2020; AlDayel and Magdy, 2021). However, the landscape of social media communication has evolved significantly, with users increasingly expressing opinions through diverse, multi-channel formats. Consequently, Multi-modal Stance Detection (MSD) has emerged to model such posts by predicting a stance (e.g., *Favor*, *Against*, or *Neutral*) toward a given target from a multi-modal input (Liang et al., 2024; Zhang et al., 2023). In practice, most studies still rely on a mainstream paradigm: a dual-encoder architecture followed by a fused prediction head, implicitly assuming that visual and textual signals complement each other to enrich stance expressions. Yet recent empirical analyses suggest that images do not consistently improve stance prediction, and the observed gains are often attributable to easy cues such as in-image text rather than general visual semantics (Vasilakes et al., 2025). These findings highlight a key challenge for MSD: while multi-modal inputs enrich stance expressions, they also make fusion harder—models must reliably assess each modality's utility and combine them robustly when their contributions are inconsistent, redundant, or potentially misleading.

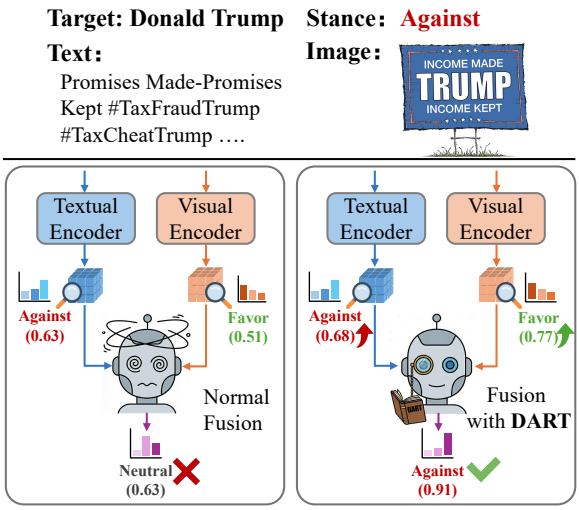

Figure 1: An example of MDD. The gold label, uni-modal predictions, and confidence scores are the **actual** test-set outputs produced by the evaluated model.

To cope with unreliable multi-modal fusion, prior work in MSD and related affective tasks has explored three common directions: (i) explicit con-

flict detection relying on manual annotations to localize cross-modal conflict (Huang et al., 2024; Shao et al., 2024); (ii) inference-time routing, e.g., dynamic routers or Mixture-of-Experts designs that select experts conditioned on the input (Yu et al., 2024; Chen et al., 2025); and (iii) data-level augmentation that constructs synthetic pairs or injects cross-modal noise to regularize representations (Wen et al., 2023; Zeng et al., 2022).

However, beyond input-level heterogeneity, these methods often overlook a subtler yet pervasive failure mode occurring *inside* the model's decision process: even when input pairs convey a semantically unified stance, the modality-specific encoders within a standard MSD model can still yield *inconsistent* stance decisions, as illustrated in Figure 1. We term this phenomenon **Modal Decision Disagreement (MDD)**. Critically, standard training objectives supervise only the final fused prediction, leaving *internal decision consistency* between the modality-specific branches unconstrained. When MDD occurs, the fusion module, without explicit guidance, can struggle to effectively arbitrate between conflicting high-confidence signals, leading to an incorrect or compromised prediction (e.g., predicting *Neutral* when one branch supports *Favor* and the other supports *Against*).

These observations suggest that improving MSD requires more than handling input-level heterogeneity: the fusion module must be trained to behave robustly when modality-specific branches disagree, rather than being optimized only through the final fused supervision. Motivated by this model-internal brittleness, we propose **DART** (**D**isagreement-**A**ware **R**obust **T**raining), a disagreement-aware training framework that strengthens fusion against internal decision inconsistency by leveraging branch-level decision signals, without any additional annotations. Specifically, we introduces a training-time auxiliary arbitration head that monitors modality-specific branch predictions, quantifies the severity of their disagreement, and injects this disagreement signal into the model's *decision representations* as an explicit training constraint, thereby directly regularizing how the fused predictor is guided under disagreement. To further reinforce this arbitration ability, we applies a text stance-flip perturbation that synthesizes counterfactual samples with deliberate cross-modal stance conflicts, exposing the arbitration objective to harder disagreement cases during training. Moreover, motivated by the observed over-reliance on in-image text, we incorporate a lightweight visual decomposition to separate embedded text cues from non-text visual content before fusion. Together, these designs address the fusion brittleness induced by MDD by making disagreement *visible at the decision level* during training and teaching the model to arbitrate conflicting high-confidence signals more reliably, rather than drifting to overly cautious compromise predictions.

Our contributions are summarized as follows:

(1) We formalize **MDD** as a decision-level, model-internal disagreement state for diagnosing fusion brittleness in MSD.

(2) We propose **DART**, a training-only arbitration framework that improves fusion under MDD without inference-time routing.

(3) Experiments on five MSD benchmarks show consistent gains across in-target and zero-shot settings, notably in MDD-conditioned analysis.

2 Related Work

Multi-modal Stance Detection. Multi-modal stance detection (MSD) extends stance classification to paired text–image posts on social media (Baltrušaitis et al., 2019; Chen et al., 2022). To support systematic evaluation, recent work has released multiple MSD datasets and benchmarks across domains (e.g., vaccination (Weinzierl and Harabagiu, 2023), climate videos (Wang et al., 2024), and multi-turn multi-modal conversations (Niu et al., 2024)). Methodologically, mainstream MSD models largely follow dual-stream encoders with cross-modal alignment/fusion, including target-oriented interaction or weighting designs, as summarized in multi-modal fusion surveys (Zhao et al., 2024; Li and Tang, 2024).

Multi-modal Inconsistency and Disagreement. At the data level, cross-modal incongruity is often treated as a semantic *signal* in affective tasks (notably multi-modal sarcasm), where models explicitly capture text–image mismatch patterns (Cai et al., 2019; Wen et al., 2023; Farabi et al., 2024). More broadly, multi-modal learning and sentiment/affect modeling often treat inconsistency as a missing/noisy/unreliable modality, motivating robustness objectives such as decomposition and missing-modality learning (Lai et al., 2023; Wu et al., 2024; Zeng et al., 2022). In parallel, inconsistency can be formulated as a detect/ground problem (e.g., contradiction or manipulation localization), typically relying on extra supervision

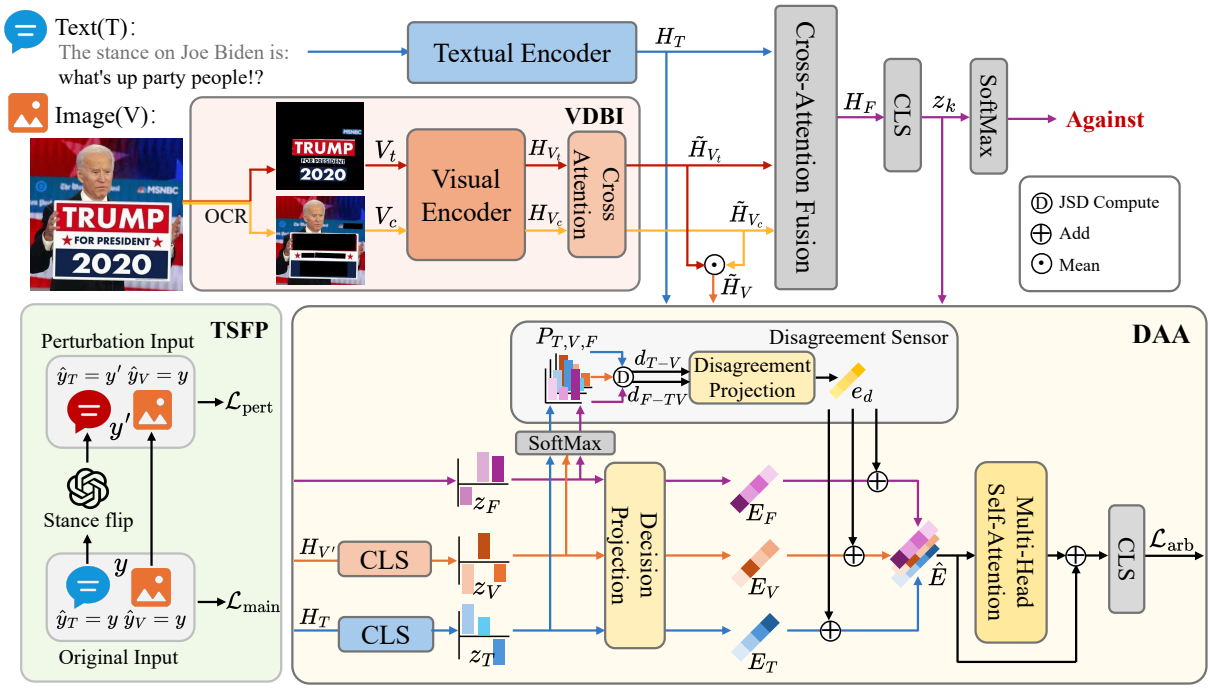

Figure 2: Overall architecture of DART for multi-modal stance detection. The model decomposes visual cues and performs multi-modal fusion, while training introduces disagreement-aware regularization to improve fusion robustness under unimodal decision disagreement.

or multi-stage designs (Shao et al., 2024; Huang et al., 2024). At the model level, routing mechanisms (e.g., interaction-expert routing or multi-modal MoE) further adapt computation to heterogeneous signals but introduce additional architectural and/or inference-time complexity (Yu et al., 2024; Chen et al., 2025; Fedus et al., 2021).

3 Methodology

This section presents our approach to MSD under MDD, with an overview of the framework shown in Figure 2. We begin by formalizing MDD as a decision-level phenomenon, and then introduce the proposed disagreement-aware fusion components together with the corresponding training objectives.

3.1 Preliminaries

Given a text–image pair $x = (x_T, x_V)$ with label $y \in \mathcal{Y}$, MSD predicts y from x. A multi-modal backbone produces modality representations (H_T, H_V) and a fused representation H_F, yielding fused logits $z_F = g_F(H_F) \in \mathbb{R}^{|\mathcal{Y}|}$ and prediction $\hat{y}_F(x) = \arg\max_{c\in\mathcal{Y}} z_F^{(c)}(x)$.

3.2 Modal Decision Disagreement

To make branch-wise decision states explicit, we attach lightweight probe heads to (H_T, H_V) for readout and block gradients to the backbone via stop-gradient: $z_k = g_k(\mathrm{sg}(H_k))$ and $\hat{y}_k(x) = \arg\max_{c\in\mathcal{Y}} z_k^{(c)}(x)$ for $k \in \{T, V\}$. We define *Modal Decision Disagreement (MDD)* as the state where the two branch decisions differ:

$$\Omega_{\mathrm{D}}(\theta) \triangleq \{x \mid \hat{y}_T(x;\theta) \neq \hat{y}_V(x;\theta)\}, \quad (1)$$

where θ denotes the current model parameters, so membership in $\Omega_{\mathrm{D}}(\theta)$ is model-defined and varies over training. In our method, we do not act on the hard event $\mathbb{I}[x \in \Omega_{\mathrm{D}}(\theta)]$; instead, we later quantify disagreement continuously from predictive distributions to support differentiable arbitration.

3.3 Visual Decomposition and Fusion

Social media images often contain heterogeneous stance cues, including embedded text (e.g., slogans, screenshots) and non-textual content (e.g., objects, scenes, symbols). Encoding an image into a single representation may entangle these cues and obscure their (dis)agreement with the text modality. We therefore introduce *Visual Decomposition with Bidirectional Interaction* (VDBI) to split the image into two cue-specific views with lightweight cross-view exchange, and then fuse text with both views via modality-level text-conditioned attention.

Decomposition with Bidirectional Interaction. The visual decomposition is performed *offline* in the data pipeline. Given the original image x_V, we apply *offline* PaddleOCR (PaddlePaddle Authors, 2020) in the data pipeline to localize text regions and construct two derived views: a text-focused view x_{V_t} that preserves detected text regions while masking the remaining background, and a content-focused view x_{V_c} that masks detected text regions while preserving the remaining visual content. A shared visual encoder $f_V(\cdot)$ encodes the two views into `[CLS]` vectors $H_{V_t} = f_V(x_{V_t})$ and $H_{V_c} = f_V(x_{V_c})$. To exchange complementary cues between the two views with minimal overhead, we perform a lightweight bidirectional cross-view attention on these vectors:

$$\tilde{H}_{V_t} = \mathrm{Attn}(H_{V_t}, H_{V_c}, H_{V_c}), \tag{2}$$

$$\tilde{H}_{V_c} = \mathrm{Attn}(H_{V_c}, H_{V_t}, H_{V_t}), \tag{3}$$

where $\mathrm{Attn}(Q, K, V)$ denotes multi-head attention with query Q, key K, and value V applied at the representation level.

Text-conditioned fusion. We fuse text with the enhanced visual views by text-conditioned attention, using H_T as the query and the stacked set $S = [H_T; \tilde{H}_{V_t}; \tilde{H}_{V_c}]$ as key/value:

$$H_F = \mathrm{Attn}(H_T, S, S), \tag{4}$$

and obtain fused logits $z_F = g_F(H_F)$ for the main MSD prediction.

Representative visual branch. To align the text branch with a single visual branch for downstream disagreement sensing, we form a representative visual representation by *view averaging*:

$$\tilde{H}_V \triangleq \tfrac{1}{2}\big(\tilde{H}_{V_t} + \tilde{H}_{V_c}\big). \tag{5}$$

Note that $\tilde{H}_V$ is employed exclusively for extracting the visual decision state and computing disagreement signals; in contrast, the fusion mechanism integrates $\tilde{H}_{V_t}$ and $\tilde{H}_{V_c}$ as distinct inputs.

3.4 Disagreement-Aware Arbitration Head

MDD highlights internal decision-level inconsistency between text and vision branches, but acting on the hard event $\hat{y}_T \neq \hat{y}_V$ is rigid and non-differentiable. We therefore introduce a training-time *Disagreement-Aware Arbitration Head* (DAA-Head) that (i) explicitly reads out branch decision states, (ii) quantifies disagreement continuously, and (iii) injects this signal to guide deliberation. DAA-Head is used **only during training**; at inference we output the fused prediction z_F. DAA-Head takes as inputs the branch logits read out in Sec 3.2 and the fused logits $z_F = g_F(H_F)$. We denote predictive distributions by $P_k = \mathrm{softmax}(z_k)$ for $k \in \{T, V, F\}$. Stop-gradient is applied only to the uni-modal readouts, so gradients from $\mathcal{L}_{\mathrm{arb}}$ do not update uni-modal representations through (z_T, z_V), while remaining fully differentiable with respect to the fused branch through z_F.

Disagreement Signals. We quantify decision-level divergence using the Jensen–Shannon divergence (JSD) (Lin, 2002). For two distributions P and Q, let $M = \frac{1}{2}(P + Q)$; the JSD is

$$\mathrm{JSD}(P, Q) = \tfrac{1}{2}\mathrm{KL}\Big(P \,\Big\|\, M\Big) + \tfrac{1}{2}\mathrm{KL}\Big(Q \,\Big\|\, M\Big), \tag{6}$$

where $\mathrm{KL}(\cdot\|\cdot)$ denotes the Kullback–Leibler divergence. Let $P_k = \mathrm{softmax}(z_k)$ for $k \in \{T, V, F\}$. We then construct a 2-D disagreement vector $d = [d_{T\text{–}V}; d_{F\text{–}TV}] \in \mathbb{R}^2$:

$$d_{T\text{–}V} = \mathrm{JSD}(P_T, P_V) \tag{7}$$

$$d_{F\text{–}TV} = \mathrm{JSD}\big(P_F, \tfrac{1}{2}(P_T + P_V)\big), \tag{8}$$

where $d_{T\text{–}V}$ measures text–vision divergence, while $d_{F\text{–}TV}$ captures the fusion deviation from the combined uni-modal evidence.

Disagreement-aware deliberation. DAA-Head produces arbitration logits z_{arb} by deliberating over (z_T, z_V, z_F) under the disagreement signal d. We first project each branch logit into a shared hidden space via a learnable projection network $\phi(\cdot)$ (Linear($|\mathcal{Y}| \to H$) $\to$LayerNorm$\to$GELU):

$$E_i = \phi(z_i), \quad i \in \{T, V, F\}. \tag{9}$$

We also project the disagreement vector into a hidden modulation vector

$$e_d = \sigma(W_d d + b_d) \in (0, 1)^H, \tag{10}$$

and inject it additively: $\tilde{E}_i = E_i + e_d$.

We stack $\tilde{E} = [\tilde{E}_T; \tilde{E}_V; \tilde{E}_F] \in \mathbb{R}^{3\times H}$ and perform self-attention deliberation:

$$\hat{E} = \mathrm{Attn}(\tilde{E}, \tilde{E}, \tilde{E}) + \tilde{E}, \tag{11}$$

followed by mean pooling $c = \frac{1}{3}\sum_i \hat{E}_i$ and a linear output head $z_{\mathrm{arb}} = W_o c + b_o$.

3.5 Text Stance-Flip Perturbation

To expose the model to more challenging disagreement regimes during training, we introduce a *text stance-flip perturbation* (TSFP). TSFP is generated *offline*: for each instance (x_T, x_V, y), we use GPT-3.5-Turbo (prompt in Appendix C) to produce a perturbed text x'_T that expresses the *opposite* stance while preserving the topic and overall linguistic style, and flip the label accordingly to y'. We keep the image unchanged, yielding (x'_T, x_V, y').

Unlike conventional augmentation that enforces cross-modal consistency, TSFP intentionally increases the likelihood and severity of text–vision divergence, providing harder training conditions for learning robust arbitration under disagreement. TSFP incurs no inference-time overhead.

3.6 Optimization

We optimize the model with three training objectives. For an original sample (x_T, x_V, y), we train the fused branch with

$$\mathcal{L}_{\text{main}} = \text{CE}(z_F, y). \tag{12}$$

We additionally train the Disagreement-Aware Arbitration Head with

$$\mathcal{L}_{\text{arb}} = \text{CE}(z_{arb}, y). \tag{13}$$

For a stance-flipped perturbed sample (x'_T, x_V, y'), we apply the same fused-branch classification loss and keep it separate from $\mathcal{L}_{\text{main}}$ to control the perturbation strength.

$$\mathcal{L}_{\text{pert}} = \text{CE}(z'_F, y'). \tag{14}$$

The overall objective is

$$\mathcal{L} = \mathcal{L}_{\text{main}} + \mathcal{L}_{\text{arb}} + \lambda_{\text{pert}}\mathcal{L}_{\text{pert}}, \tag{15}$$

where λ_{pert} weight the perturbation loss.

4 Experiments

4.1 Experimental Setup

Datasets. We evaluate on five multi-modal stance detection datasets (MTSE, MCCQ, MWTWT, MRUC, and MTWQ) from the benchmark of Liang et al. (Liang et al., 2024). They consist of Twitter image–text pairs spanning diverse domains (e.g., US elections, COVID-19, and geopolitical conflicts). We follow the official splits and evaluation protocols, and report macro-F1 (%).

Baselines. We compare against representative baselines in three groups: **Text-only** (BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), KEBERT (Kawintiranon and Singh, 2022), LLaMA2-70B (Touvron et al., 2023), GPT-4), **Visual-only** (ResNet-50 (He et al., 2016), ViT-Base (Dosovitskiy et al., 2021), SwinV2-Base (Liu et al., 2021)), and **Multi-modal** (RoBERTa+ViT late fusion, ViLT (Kim et al., 2021), CLIP (Radford et al., 2021), Qwen-VL (Bai et al., 2023), GPT-4V, and TMPT (Liang et al., 2024) as the primary state-of-the-art baseline). We additionally report TMPT+CoT (Liang et al., 2024), which augments inference with GPT-4V chain-of-thought prompting. For a strict **no-test-augmentation** comparison, we use a **training-only** variant TMPT+CoT$^\sharp$ where CoT augmentation is applied during training but disabled at inference.

Implementation details. We use RoBERTa-base and ViT-base/224 as backbones. We train in PyTorch with AdamW for 25 epochs (batch size 22, weight decay 0.02), using learning rates $1\text{e}{-5}$ for encoders and $1\text{e}{-4}$ for task heads with 10% warmup. We use a unified hyperparameter setting across all datasets; in particular, $\lambda_{\text{pert}} = 0.5$. All experiments are conducted on a single NVIDIA RTX 4090 GPU. Following common practice in MSD, we fix the random seed for all experiments to ensure reproducibility.

4.2 Main Results

Overview. Table 1 and Table 2 report the in-target and zero-shot results (Macro-F1, %) on the multi-modal stance detection benchmark. DART consistently improves over prior trainable multi-modal baselines that do not rely on test-time augmentation. Results for large proprietary models (e.g., GPT-4V) are included for reference, but they are not strictly comparable to fully-trainable methods due to differences in training data, capacity, and inference-time tools.

In-target performance. In Table 1, **DART** outperforms TMPT on *all* reported targets, indicating robust in-domain improvements rather than gains limited to a small subset. Compared with the RoBERTa+ViT backbone baseline, DART is also consistently stronger across targets, underscoring the need for principled fusion beyond straightforward backbone pairing. As one example, on MCCQ, DART achieves a relative gain of 15.0% over TMPT. While TMPT+CoT reports higher

Modality	Method	MTSE		MCCQ	MWTWT					MRUC		MTWQ	
		DT	JB	CQ	CA	CE	AC	AH	DF	RUS	UKR	MOC	TOC
Textual	BERT	52.20	52.25	63.02	74.26	51.71	59.34	53.43	77.90	41.25	42.04	54.13	45.91
	RoBERTa	57.33	59.51	69.30	74.61	54.63	63.26	51.67	79.21	41.68	52.67	57.37	52.82
	KEBERT	64.77	69.24	71.02	74.06	54.97	65.78	61.16	80.57	39.69	58.80	59.31	45.81
	LLaMA2	53.23†	52.67†	47.40†	34.89†	41.95†	49.09†	44.32†	30.21†	38.84†	38.54†	55.31†	46.51†
	GPT-4	68.74†	66.39†	65.84†	63.14†	65.12†	69.93†	71.62†	52.69†	41.64†	53.76†	58.05†	49.81
Visual	ResNet	39.31	40.10	47.90	39.91	41.71	43.08	38.64	51.07	34.59	40.51	40.43	34.16
	ViT	35.45	40.35	47.18	41.18	42.28	45.49	38.82	54.89	34.02	38.52	38.76	36.27
	SwinT	38.90	40.44	46.26	39.29	43.55	45.49	42.88	54.52	34.39	42.14	36.20	36.10
Multi-modal	RoBERTa+ViT	41.86	45.82	61.32	63.20	44.71	56.45	46.85	73.71	39.28	48.41	47.47	40.86
	ViLT	39.83	47.13	46.33	58.31	48.13	53.90	43.09	68.68	34.68	41.69	44.93	42.02
	CLIP	57.72	65.25	62.74	65.71	48.58	62.26	56.77	79.55	40.66	47.49	60.00	36.89
	Qwen-VL	43.31†	45.13†	50.51†	43.06†	45.49†	49.79†	46.04†	27.73†	36.50†	40.78†	42.14†	39.34†
	GPT-4V	**70.46**†	**72.82**†	61.63†	44.59†	47.07†	57.47†	57.90†	37.61†	44.83†	56.40†	66.72†	56.90†
	TMPT	55.41†	61.61†	67.67†	76.60†	63.19†	67.25†	62.92†	81.19†	43.56†	59.24†	55.68†	46.82†
	TMPT+CoT	66.61†	68.75†	71.79†	74.40†	69.96†	68.43†	63.00†	82.71†	45.04†	60.52†	**68.95**†	**59.87**†
	TMPT+CoT♯	50.85	57.89	71.79	74.31	64.87	69.73	62.24	82.11	42.65	58.29	59.59	50.77
	DART (Ours)	61.17	68.11	**77.83**	**80.45**	**74.45**	**71.65**	**70.49**	82.63	**49.40**	**64.13**	62.57	54.63
	- w/o VDBI	60.39	67.81	75.65	77.82	73.79	70.69	69.25	82.45	49.10	63.21	62.17	54.53
	- w/o TSFP	60.44	65.46	72.12	79.13	73.10	71.33	69.09	82.50	45.07	56.13	61.16	47.89
	- w/o DAA	58.87	60.34	71.63	76.86	72.39	70.01	67.95	81.17	41.86	55.80	59.28	52.37
	- VDBI only	57.08	58.40	68.91	77.83	71.04	66.68	66.65	82.05	41.35	57.44	55.48	42.58
	- TSFP only	59.52	61.73	66.91	73.48	70.55	66.18	64.61	81.04	39.09	49.48	59.06	47.63
	- DAA only	57.62	62.61	70.29	76.56	62.81	69.06	70.15	**83.61**	44.37	56.71	59.40	52.40

Table 1: Experimental results (% Macro-F1) for in-target multi-modal stance detection. † from (Liang et al., 2024). ♯ denotes our training-only variant of TMPT+CoT, where GPT-4V CoT is used only for training-time augmentation and disabled at inference. The shaded row indicates our proposed **DART**; rows beneath it report ablations (*w/o* a component) and component-only variants. Bold values denote the best result in each column.

Modality	Method	MTSE		MWTWT				MRUC		MTWQ	
		DT	JB	CA	CE	AC	AH	RUS	UKR	MOC	TOC
Textual	BERT	28.17	29.93	58.04	56.98	48.75	44.15	22.01	15.45	28.04	9.57
	RoBERTa	29.94	34.21	71.53	70.80	72.16	62.27	24.23	24.25	25.42	11.53
	KEBERT	24.92	35.22	61.60	62.17	67.17	57.03	24.18	30.06	29.64	17.50
	LLaMA2	53.57†	53.92†	32.47†	38.37†	48.08†	46.13†	31.86†	36.34†	51.46†	44.10†
	GPT-4	70.78†	68.83†	57.19†	60.56†	65.63†	69.01†	40.22†	49.18†	62.10†	52.12†
Visual	ResNet	26.09	29.60	23.05	24.64	26.06	26.03	23.40	25.96	27.61	24.96
	ViT	25.28	28.39	22.62	23.81	29.02	26.38	26.17	28.48	29.37	23.69
	Swin-T	26.78	27.10	23.56	25.32	31.89	26.13	25.43	24.33	27.91	19.73
Multi-modal	RoBERTa+ViT	26.70	31.57	59.21	59.30	65.04	59.28	23.33	15.21	24.76	11.70
	ViLT	28.08	29.74	38.33	46.00	55.01	48.55	21.56	23.96	23.54	19.18
	CLIP	28.21	28.99	61.08	55.67	63.80	60.06	25.62	27.40	27.21	15.69
	Qwen-VL	47.62†	46.14†	38.57†	43.36†	47.82†	41.01†	36.95†	41.39†	44.32†	44.08†
	GPT-4V	**72.68**†	**71.28**†	42.23†	45.92†	54.59†	53.19†	42.09†	47.00†	**65.00**†	**52.36**†
	TMPT	31.69†	32.65†	66.36†	66.39†	66.32†	61.56†	23.87†	24.71†	32.18†	26.48†
	TMPT+CoT	54.30†	58.46†	67.28†	63.73†	64.87†	54.26†	**48.99**†	**51.75**†	45.32†	43.70†
	TMPT+CoT♯	37.14	35.15	70.25	72.18	72.19	56.08	25.65	25.20	28.71	27.46
	DART (Ours)	37.50	40.01	**75.10**	**73.42**	**75.34**	**69.39**	32.83	42.89	41.69	31.55
	- w/o VDBI	33.59	38.79	74.48	73.59	74.13	67.44	23.23	24.27	34.84	24.84
	- w/o TSFP	36.76	38.80	73.01	72.27	73.40	68.97	25.17	25.14	35.52	27.23
	- w/o DAA	31.75	38.56	73.16	72.18	70.76	68.24	26.06	26.18	31.32	27.38
	- VDBI only	33.48	34.08	71.02	71.83	73.30	65.04	24.35	25.87	35.40	29.51
	- TSFP only	31.59	33.89	72.04	72.78	73.22	65.98	23.18	24.51	35.43	28.21
	- DAA only	33.91	38.79	73.04	72.34	73.64	65.88	24.12	24.55	32.02	22.13

Table 2: Experimental results (% Macro-F1) for zero-shot multi-modal stance detection. † from (Liang et al., 2024). ♯ denotes our training-only variant of TMPT+CoT, where GPT-4V CoT is used only for training-time augmentation and disabled at inference. The shaded row indicates our proposed **DART**; rows beneath it report ablations (*w/o* a component) and component-only variants. Bold values denote the best result in each column.

MTWQ scores, it relies on test-time CoT; under the training-only TMPT+CoT$^\sharp$, DART remains higher on both MTWQ targets, suggesting the gains persist without inference-time augmentation.

Zero-shot performance. In Table 2, **DART** consistently surpasses TMPT across all reported targets under target shift, and it also improves over the RoBERTa+ViT backbone baseline, indicating stronger transfer beyond a backbone-only effect. These gains are most pronounced on the harder-to-generalize targets such as MRUC, while remaining uniformly positive on MWTWT. For example, on MRUC-UKR, DART achieves a 73.6% relative gain over TMPT. Meanwhile, TMPT+CoT can be stronger on some MTSE/MRUC targets, however, DART remains competitive and provides a stronger fully-trainable alternative under the same evaluation protocol. Moreover, the TMPT+CoT$^\sharp$ shows only marginal improvements under target shift, suggesting that using CoT solely as training-time augmentation provides limited transfer benefits.

4.3 Ablation Study

Table 1 and Table 2 show that DART's improvements are driven by *complementary* mechanisms rather than loosely stacked add-ons: removing any component degrades performance, and the degradation is generally more severe in the zero-shot setting, consistent with our claim that DART targets robustness under target shift. Among the three modules, DAA is the key supervisory signal that turns modality-level disagreement into actionable training guidance; correspondingly, removing DAA causes the most salient collapses on disagreement-sensitive targets, for example reducing zero-shot MTWQ-MOC by 10.37 points. At the same time, VDBI and TSFP are not redundant auxiliaries: they improve the conditions under which arbitration can be learned by stabilizing visual evidence and broadening exposure to hard disagreement regimes, respectively, and their removal particularly hurts transfer performance; notably, removing VDBI drops zero-shot MRUC-UKR by 18.62 points. Finally, the component-only variants remain clearly below the full model across the benchmark, occasionally matching or exceeding the full system on isolated targets but failing to deliver consistent gains, which supports our story line that DAA provides the arbitration objective while VDBI/TSFP supply the reliable and diverse disagreement contexts needed for it to generalize.

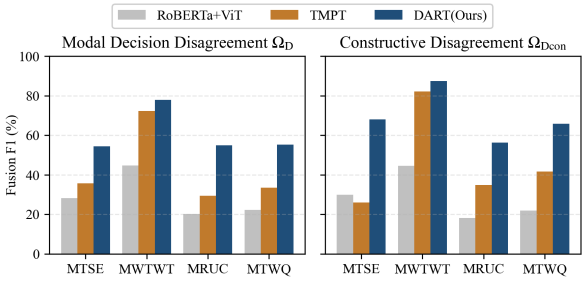

Figure 3: Zero-shot fusion Macro-F1 (%) on disagreement slices Ω_{D} and Ω_{Dcon} Slices are constructed using lightweight stop-gradient probes to read out uni-modal decisions without altering the backbone behavior.

4.4 Disagreement-focused Analysis

Tables 1–2 aggregate (dis)agreement-dominant instances. Since DART targets stabilization under MDD, we further analyze model behavior on $\Omega_{\mathrm{D}}(\theta)$ slices to ensure the gains are not dominated by agreement-heavy cases. We additionally consider a constructive MDD slice where exactly one uni-modal branch is correct:

$$\delta(x;\theta) \triangleq (\hat{y}_T(x;\theta) = y) \oplus (\hat{y}_V(x;\theta) = y), \tag{16}$$

$$\Omega_{\mathrm{Dcon}}(\theta) \triangleq \{x \in \Omega_{\mathrm{D}}(\theta) \mid \delta(x;\theta)\}. \tag{17}$$

Figure 3 shows that the standard late-fusion baseline (RoBERTa+ViT) performs poorly under $\Omega_{\mathrm{D}}(\theta)$, suggesting that label-only fusion is fragile when branch-wise decisions disagree. TMPT, while stronger overall, still weakens noticeably in this MDD state. In contrast, DART improves fusion Macro-F1 across all evaluated targets on $\Omega_{\mathrm{D}}(\theta)$, with typically larger gains on $\Omega_{\mathrm{Dcon}}(\theta)$ where exactly one uni-modal branch is correct in hindsight. Moreover, the gains are typically larger on the analysis-only slice $\Omega_{\mathrm{Dcon}(\theta)}$, which isolates disagreement cases where exactly one branch decision matches the gold label in hindsight. Together, these results suggest that DART more effectively stabilizes fusion under branch disagreement and that its improvements are not driven solely by agreement-dominant instances. Note that $\Omega_{\mathrm{D}}(\theta)$ and $\Omega_{\mathrm{Dcon}}(\theta)$ are model-dependent internal decision states (rather than fixed data partitions), so slice membership can vary across methods; we therefore report slice coverage and a controlled comparison under the same RoBERTa+ViT encoders in Appendix B.

Figure 4 suggests an association between fusion deviation and predictive correctness, and this association becomes more pronounced on the disagreement slice $\Omega_{\mathrm{D}}(\theta)$. Compared with the

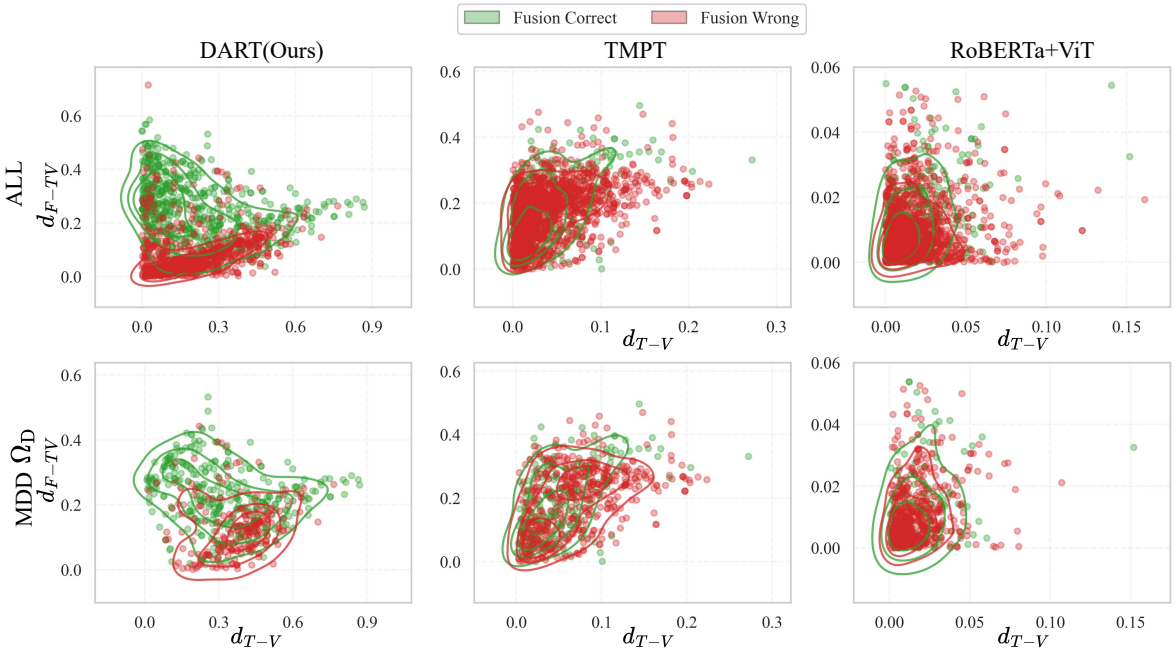

Figure 4: Visualization of fusion behavior on MTWQ in the zero-shot setting (target: TOC), shown for **all test instances** (top row) and the **MDD slice** Ω_{D} (bottom row). We compare **DART** (Ours), **TMPT**, and the **RoBERTa+ViT** late-fusion baseline. Each point is a test instance plotted by text–vision divergence d_{T-V} versus fusion deviation d_{F-TV}. Contours indicate KDE density for correct (green) and incorrect (red) predictions. **Axis limits are independently scaled across columns for readability; comparisons focus on within-panel geometry.**

full set, $\Omega_{\mathrm{D}}(\theta)$ focuses on instances with branch-level decision disagreement, which makes the contrast between correct and incorrect predictions along d_{F-TV} more visually salient. Intuitively, this regime reduces the influence of agreement-dominant instances and stresses whether the fusion mechanism can reliably arbitrate conflicting unimodal signals when the two branches disagree. Within its panel, the RoBERTa+ViT late-fusion baseline clusters toward the low-d_{T-V} region and shows limited variation along d_{F-TV}, indicating that its fused prediction tends to remain close to the mean of uni-modal decisions across the observed disagreement cases. TMPT occupies a wider portion of the d_{T-V} axis within its panel, but the KDE contours of correct and incorrect predictions still overlap heavily (notably within $\Omega_{\mathrm{D}}(\theta)$), suggesting limited separability along the fusion-deviation axis under disagreement. In contrast, DART exhibits a clearer *vertical stratification of correctness*: correct predictions concentrate at higher d_{F-TV}, while errors are skewed toward lower d_{F-TV}, and this stratification is stronger on $\Omega_{\mathrm{D}}(\theta)$ than on the full set. Since d_{F-TV} measures how far the fused distribution deviates from the mean of uni-modal decisions, the observed stratification is consistent with our training-time arbitration signal shaping fusion under MDD, making fusion deviation more predictive of correctness when the modalities disagree. This pattern also aligns with the intended role of arbitration: enabling confident departures from the unimodal mean when one modality provides the decisive evidence, rather than averaging away informative disagreement. We additionally provide a qualitative case study to illustrate representative failure modes and how DART resolves branch disagreement in Appendix A.

5 Conclusion

In this work, we uncover that model-internal inconsistency is a critical bottleneck in multi-modal fusion. We formalize this phenomenon as MDD, serving as an instructive state to diagnose fusion brittleness. Building on this insight, our proposed **DART** framework teaches the model to arbitrate conflicting signals during training, keeping inference efficient. Empirical results demonstrate that DART significantly improves robustness, particularly on samples with high disagreement. Beyond performance gains, our findings suggest that being *aware* of such decision-level conflicts enables models to better grasp the implicit, often subtle nuances in social media expressions, paving the way for more reliable multi-modal understanding.

Limitations

Our analyses and disagreement-conditioned evaluations depend on auxiliary unimodal heads to derive text-only and image-only decisions, so measured disagreement rates may vary with head design; the OCR-based visual decomposition is sensitive to OCR errors and configuration; and the LLM-based stance-flip perturbation may introduce label noise and reproducibility issues due to non-deterministic generation. For fair comparison, we instantiate DART primarily on a RoBERTa+ViT backbone with a specific fusion design, and a systematic study across stronger/multilingual encoders and alternative fusion operators is left for future work.

Ethical Considerations

Our study relies exclusively on publicly available social media datasets. Nevertheless, analyzing multimodal social media content for stance detection may raise concerns related to user privacy and responsible data usage. Our method does not require access to any private, sensitive, or personally identifiable information. All social media posts are processed in anonymized form, and no user identities, metadata, or personal attributes are retained or exposed, ensuring that the data cannot be traced back to individual users. The task of multi-modal stance detection has non-trivial social implications. Prediction errors—especially under modal disagreement—may mislabel user intent and mislead downstream decisions. We therefore emphasize robustness-focused training and recommend cautious deployment as a decision-support tool rather than an autonomous decision-maker. Finally, we commit to responsible research practices in code and model dissemination. Any released implementation will adhere to established ethical guidelines, clearly document intended use cases and limitations, and discourage misuse in high-stakes or privacy-sensitive applications.

References

Abeer AlDayel and Walid Magdy. 2021. Stance detection on social media: State-of-the-art and trends. *Information Processing & Management*, 58(4):102597.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *Preprint*, arXiv:2308.12966.

Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. 2019. Multimodal machine learning: A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(2):423–443.

Yitao Cai, Huiyu Cai, and Xiaojun Wan. 2019. Multi-modal sarcasm detection in twitter with hierarchical fusion model. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2506–2515, Florence, Italy. Association for Computational Linguistics.

Kezhou Chen, Shuo Wang, Huixia Ben, Shengeng Tang, and Yanbin Hao. 2025. Mixture of multimodal adapters for sentiment analysis. 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 1822–1833.

Shizhe Chen, Qin Zou, and Xiaodong Chen. 2022. Vision-language pre-training: Basics, recent advances, and future trends. *Preprint*, arXiv:2202.09061.

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, Minneapolis, Minnesota. Association for Computational Linguistics.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations (ICLR)*.

Shafkat Farabi, Tharindu Ranasinghe, Diptesh Kanojia, Yu Kong, and Marcos Zampieri. 2024. A survey of multimodal sarcasm detection. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence (IJCAI-24)*, pages 8020–8028.

William Fedus, Barret Zoph, and Noam Shazeer. 2021. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Preprint*, arXiv:2101.03961.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778.

Mingzhen Huang, Shan Jia, Zhou Zhou, Yan Ju, Jialing Cai, and Siwei Lyu. 2024. Exposing text-image inconsistency using diffusion models. In *The Twelfth International Conference on Learning Representations (ICLR)*.

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 (LREC 2022)*, pages 7360–7367, Marseille, France. European Language Resources Association.

Wonjae Kim, Bokyung Son, and Ildoo Kim. 2021. Vilt: Vision-and-language transformer without convolution or region supervision. *Preprint*, arXiv:2102.03334.

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

Xiaoxi Lai, Shuai Wang, Ruiqiang Xiong, Yanjun Mao, Yanbin Hao, Wen Han, and Baochang Zhang. 2023. Multimodal sentiment analysis: A survey. *Preprint*, arXiv:2305.07611.

Songtao Li and Hao Tang. 2024. Multimodal alignment and fusion: A survey. *Preprint*, arXiv:2411.17040.

Bin Liang and 1 others. 2024. Multi-modal stance detection: New benchmarks and model. In *ACL*.

Jianhua Lin. 2002. Divergence measures based on the shannon entropy. *IEEE Transactions on Information theory*, 37(1):145–151.

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

Ze Liu, Han Hu, Yutong Lin, Zhuliang Yao, Zhenda Xie, Yixuan Wei, Jia Ning, Yue Cao, Zheng Zhang, Li Dong, Furu Wei, and Baining Guo. 2021. Swin transformer v2: Scaling up capacity and resolution. *Preprint*, arXiv:2111.09883.

Zihan Niu and 1 others. 2024. Multimodal multi-turn conversation stance detection: A challenge dataset and effective model. *Preprint*, arXiv:2409.00597.

PaddlePaddle Authors. 2020. Paddleocr. `https://github.com/PaddlePaddle/PaddleOCR`. GitHub repository, accessed 2026-01-02.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. *Preprint*, arXiv:2103.00020.

Rui Shao, Tianxing Wu, Jianlong Wu, Liqiang Nie, and Ziwei Liu. 2024. Detecting and grounding multimodal media manipulation and beyond. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(8):5556–5574.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, and 49 others. 2023. Llama 2: Open foundation and fine-tuned chat models. *Preprint*, arXiv:2307.09288.

Jake Vasilakes, Carolina Scarton, and Zhixue Zhao. 2025. Exploring vision language models for multimodal and multilingual stance detection. *arXiv preprint arXiv:2501.17654*.

Jiawen Wang, Longfei Zuo, Siyao Peng, and Barbara Plank. 2024. Multiclimate: Multimodal stance detection on climate change videos. In *Proceedings of the Third Workshop on NLP for Positive Impact*, pages 315–326, Miami, Florida, USA. Association for Computational Linguistics.

Markus A. Weinzierl and Sanda M. Harabagiu. 2023. MMVax: Multimodal stance detection for vaccination-related Twitter conversations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12640–12655, Singapore. Association for Computational Linguistics.

Changsong Wen, Guoli Jia, and Jufeng Yang. 2023. Dip: Dual incongruity perceiving network for sarcasm detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2540–2550.

Renjie Wu, Hu Wang, Hsiang-Ting Chen, and Gustavo Carneiro. 2024. Deep multimodal learning with missing modality: A survey. *Preprint*, arXiv:2409.07825.

Haofei Yu, Zhengyang Qi, Lawrence Keunho Jang, Russ Salakhutdinov, Louis-Philippe Morency, and Paul Pu Liang. 2024. MMoE: Enhancing multimodal models with mixtures of multimodal interaction experts. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 10006–10030, Miami, Florida, USA. Association for Computational Linguistics.

Jiandian Zeng, Jiantao Zhou, and Tianyi Liu. 2022. Mitigating inconsistencies in multimodal sentiment analysis under uncertain missing modalities. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2924–2934, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Qiang Zhang and 1 others. 2023. Investigating the robustness of multimodal stance detection models. In *EMNLP*.

Fei Zhao, Chengcui Zhang, and Baocheng Geng. 2024. Deep multimodal data fusion. *ACM Computing Surveys*, 56(9):1–36.

Text: nytimes Brother, can you spare ten trillion dimes?? #TrumpTaxReturns #TrumpIsBroke #Debates2020 #TrumpTaxes #trump #BREAKING TrumpLiedPeopleDied #TrumpFailed #MondayMotivation #MondayMorning

Image:

Model	T	V	F
RoBERTa+ViT	Favor(0.353)×	Neutral(0.473)×	Favor(0.391)×
TMPT	Favor(0.472)×	Against(0.519)✓	Favor(0.517)×
DART	Favor(0.499)×	Against(0.515)✓	Against(0.700)✓

Figure 5: Case 1 from Ω_{Dcon} on MTSE-DT (zero-shot), gold: AGAINST.

Text: Promises Made-Promises Kept #TaxFraudTrump #TaxCheatTrump #TaxFraud #PromisesMadePromisesKept #BusinessFraud #Debates2020 #TrumpIncomeTaxes #TrumpDebt

Image:

Model	T	V	F
RoBERTa+ViT	Favor(0.376)×	Favor(0.508)×	Favor(0.424)×
TMPT	Against(0.634)✓	Favor(0.514)×	Neutral(0.635)×
DART	Against(0.683)✓	Favor(0.775)×	Against(0.908)✓

Figure 6: Case 2 from Ω_{Dcon} on MTSE-DT (zero-shot), gold: AGAINST.

A Case Study

We present two representative zero-shot MTSE-DT instances from Ω_{Dcon} (Section 3.2), i.e., *constructive disagreement* cases where the unimodal text and image decisions disagree and exactly one modality matches the gold label (AGAINST). For transparency, the table under each example reports the predicted class and confidence for the attached text-only head (T), image-only head (V), and the fused prediction (F). The unimodal heads are used for analysis only, and we stop gradients before the encoders to avoid altering the backbone representations. Figure 5 illustrates a case where the text modality is misleading: both TMPT and DART assign FAVOR to T, while V predicts AGAINST. DART correctly arbitrates toward the visual evidence and outputs AGAINST with high confidence (0.700). In contrast, TMPT follows the misleading textual cue and predicts FAVOR (0.517), and the late-fusion baseline is also incorrect. Figure 6 shows the complementary pattern where the text modality is reliable: T correctly predicts AGAINST, but V is distracted by the campaign-style image and predicts FAVOR. Under this contradiction, TMPT collapses to NEUTRAL (0.635), whereas DART remains decisive and predicts AGAINST (0.908), consistent with selecting the reliable modality when the disagreement is solvable.

Overall, these examples provide qualitative evidence consistent with DART improving fusion robustness on Ω_{Dcon}, where correct modality arbitration is feasible and materially affects the final decision.

B MDD State Coverage

To contextualize our disagreement-focused analyses, we report the **coverage** of the modal decision disagreement (MDD) state, i.e., the fraction of test instances that fall into $\Omega_{\text{D}}(\theta)$ under a given θ. We compare RoBERTa+ViT, TMPT, and DART across benchmarks in the zero-shot and in-target settings. Higher values indicate that a larger fraction of instances enters the MDD state for the corresponding model, reflecting a *model-state* characterization rather than a data-defined partition. Figures 7–8 show that the MDD state is not rare and that its coverage varies substantially across benchmarks and methods, reinforcing that disagreement is best understood as a *model-induced internal decision state* rather than an intrinsic property of the data. Moreover, MDD coverage tends to be higher and more uneven in the zero-shot setting, consistent with target shift amplifying cross-modal conflict. Across methods, the coverage profiles differ, indicating that the frequency with which models enter the MDD state is itself method-dependent; consequently, disagreement-conditioned performance comparisons should be interpreted together with these coverage statistics.

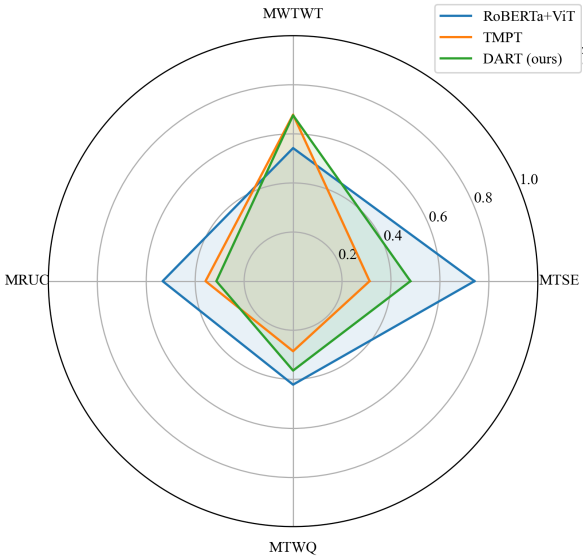

Figure 7: Zero-shot MDD Coverage.

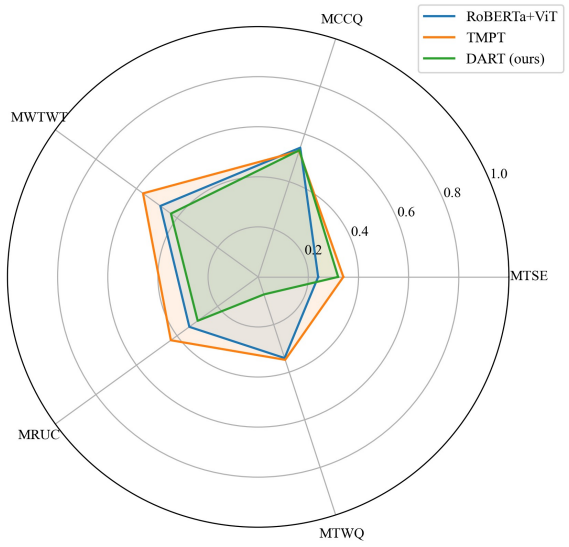

Figure 8: In-target MDD Coverage.

C Prompt Templates

This appendix reports the prompt templates used in our LLM-based rewriting pipeline for reproducibility. We present prompts as plain text (instead of images) to ensure copyability, accessibility, and faithful reproduction of the input-output contract. For typesetting, we insert line breaks; in implementation we normalize whitespace before sending prompts to the LLM.

To synthesize counterfactual textual samples with controlled stance relations, we employ an LLM to *minimally* revise the original sentence with respect to a specified target relation to the topic. We use three prompt templates corresponding to (i) stance→stance, (ii) stance→unrelated, and (iii) unrelated→stance transformations (Listings 1–3). All templates enforce a JSON-only output with a single required field (`"Revised Sentence"`), enabling deterministic parsing in our pipeline. We discard and re-query generations that violate the JSON schema. Table 3 summarizes the placeholders used to instantiate the prompts.

Placeholder	Description
`{SRC_STANCE}`	Source stance label of the input sentence (known).
`{TGT_STANCE}`	Target stance label to rewrite into.
`{SENTENCE}`	Input sentence x to be minimally edited.
`{TOPIC}`	Topic t that the stance is expressed toward.

Table 3: Variables used to instantiate the prompt templates.

C.1 Prompt Listings

Listing 1: Stance→stance rewriting prompt used in our pipeline.

```
I will give you a <Sentence> about a <Topic>, and known that
    ↪ the <Sentence>
expresses a {SRC_STANCE} attitude to the <Topic>.
Please make minimal changes to the <Sentence> so that the
    ↪ <Revised Sentence>
expresses a {TGT_STANCE} attitude to the <Topic>.
The <Sentence> is "{SENTENCE}".
The <Topic> is "{TOPIC}".
Please answer this question directly without any explanation
    ↪ with JSON format:
{
  "Revised Sentence": "Your revised sentence."
}
```

Listing 2: Stance→unrelated rewriting prompt used in our pipeline.

```
I will give you a <Sentence> about a <Topic>, and known that
    ↪ the <Sentence>
expresses a {SRC_STANCE} attitude to the <Topic>.
Please make minimal changes to the <Sentence> so that the
    ↪ <Revised Sentence>
is unrelated to the <Topic>.
The <Sentence> is "{SENTENCE}".
The <Topic> is "{TOPIC}".
Please answer this question directly without any explanation
    ↪ with JSON format:
{
  "Revised Sentence": "Your revised sentence."
}
```

Listing 3: Unrelated→stance rewriting prompt used in our pipeline.

```
I will give you a <Sentence> about a <Topic>, and known that
    ↪ the <Sentence>
is unrelated to the <Topic>.
Please make minimal changes to the <Sentence> so that the
    ↪ <Revised Sentence>
expresses a {TGT_STANCE} attitude to the <Topic>.
The <Sentence> is "{SENTENCE}".
The <Topic> is "{TOPIC}".
Please answer this question directly without any explanation
    ↪ with JSON format:
{
  "Revised Sentence": "Your revised sentence."
}
```