Information Sciences 725 (2026) 122704

Contents lists available at ScienceDirect

Information Sciences

journal homepage: www.elsevier.com/locate/ins

# 3WD-DRT: A three-way decision enhanced dynamic routing transformer for cost-sensitive multimodal sentiment analysis

Haoyu Jiang [a], Xiaoliang Chen [a, b, e, *], Duoqian Miao [b], Hongyun Zhang [b], Xiaolin Qin [c], Shangyi Du [d], Peng Lu [e]

[a] School of Computer and Software Engineering, Xihua University, Chengdu 610039, PR China
[b] College of Electronic and Information Engineering, Tongji University, Shanghai 201804, PR China
[c] Chengdu Institute of Computer Applications, Chinese Academy of Sciences, Chengdu 610041, PR China
[d] Department of Computer Science, Mcgill University, Montreal, QC, H3A 0G4, Canada
[e] Department of Computer Science and Operations Research, University of Montreal, Montreal, QC H3C3J7, Canada

ARTICLE INFO

*Keywords:*
Multimodal sentiment analysis (MSA)
Three-way decision theory
Dynamic routing transformer
Emotion-aware fusion

ABSTRACT

Accurately interpreting human emotion from language, facial expressions, and vocal tones remains a fundamental challenge in artificial intelligence. Current Multimodal Sentiment Analysis (MSA) models often struggle with two key issues. First, their static fusion strategies fail to handle conflicting modalities, such as sarcasm. Second, their standard loss functions ignore the asymmetric risks of severe misjudgments. To address these limitations, we propose the Three-Way Decision Enhanced Dynamic Routing Transformer (3WD-DRT), a framework operating on a "quality-aware, decision-driven" principle. It dynamically assesses each modality's quality using a three-way decision gate, implemented via a dedicated MLP, to partition information into acceptance, deferment, or rejection pathways. This enables the model to amplify informative signals, moderately scale uncertain ones (deferment), and attenuate noisy or misleading ones. We also introduce a novel cost-sensitive loss function that imposes greater penalties on major semantic errors, such as polarity misclassifications. This approach better aligns the model's training objective with human perception. Extensive experiments on CH-SIMS, CH-SIMSv2, MOSI, and MOSEI datasets show that 3WD-DRT consistently outperforms state-of-the-art methods, setting new benchmarks with F1-scores of 87.08 % on MOSI and 88.26 % on MOSEI. This work provides a robust solution for MSA, fostering more nuanced and reliable emotionally-aware AI systems.

## 1. Introduction

The expression of human emotion is inherently multimodal, conveyed through channels such as language, facial expressions, gestures, and vocal tone. Multimodal Sentiment Analysis (MSA) is a core branch of artificial intelligence that aims to decode and quantify complex emotions from heterogeneous data streams. By emulating the holistic way humans perceive the world, MSA has found broad applications in fields such as human-computer interaction, medical diagnosis, and public opinion monitoring. While early research emphasized discrete emotion classification, recent work has shifted toward emotion regression—predicting emotional values on a continuous scale—to better capture subtle variations in emotional intensity. A model that accurately predicts emotional polarity and intensity can both recognize basic emotions and distinguish fine-grained differences, such as varying degrees of joy. This capability is essential for developing intelligent systems with genuine empathetic capacity.

* Corresponding author at: School of Computer and Software Engineering, Xihua University, Chengdu 610039, PR China.
*Email address:* chenxl@mail.xhu.edu.cn (X. Chen).

https://doi.org/10.1016/j.ins.2025.122704
Received 27 June 2025; Received in revised form 18 September 2025; Accepted 18 September 2025
Available online 30 September 2025

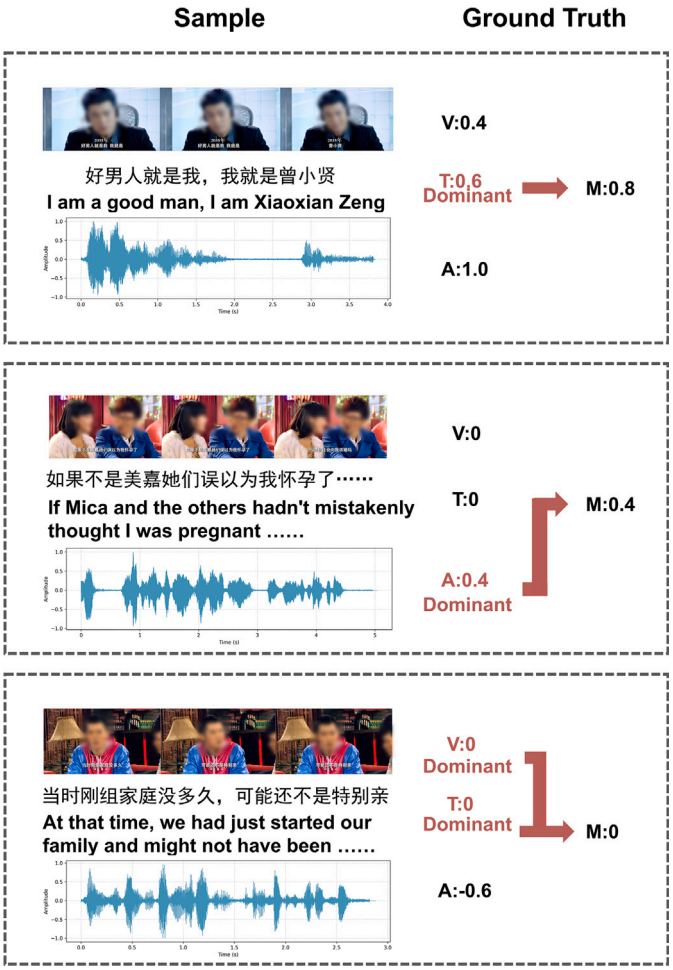


**Fig. 1.** Illustration of dynamic modal dominance with three samples from the CH-SIMS v2.0 dataset [1] (adapted and slightly blurred for anonymization). Sentiment labels for vision (V), text (T), audio (A), and the overall multimodal ground truth (M) are rated on a scale from −1 (negative) to 1 (positive). The samples show different scenarios where a single modality's sentiment score (e.g., V, T, or A) most closely matches the overall score (M), indicating a shift in modal dominance. Images are used solely for non-commercial academic research in accordance with the dataset's terms of use.

Achieving high-precision emotion regression involves three fundamental challenges. First, there is the issue of inter-modal information asymmetry and noise. In real-world interactions, the reliability of each modality is dynamic; for example, in sarcastic speech, text may mislead while vocal tone conveys the true sentiment. Robust fusion requires adaptive assessment of each modality's contribution and suppression of noisy inputs. Second, existing fusion mechanisms are often static. Common strategies, such as early feature concatenation or fixed attention-based weighting, lack the ability to adjust to sample-specific characteristics and context-dependent inter-modal dependencies. Third, traditional regression loss functions, such as Mean Squared Error (MSE), assign equal weight to all errors, disregarding the asymmetric nature of emotional misjudgment. In sentiment analysis, confusing "positive" with "negative" is far more costly than mistaking it for "neutral", yet conventional loss functions do not reflect this risk asymmetry.

Various fusion strategies have been proposed to address these challenges, but each has limitations. Tri-modality models such as MulT [2] capture deep inter-modal interactions but often assume a fixed hierarchy of modal importance, while text-centric approaches [3] employ mechanisms like sparse gating to prioritize textual data. These assumptions are problematic in cases such as sarcasm or irony, where visual or acoustic cues are more informative. Our observations indicate that modal dominance shifts dynamically. As shown in Fig. 1, in some cases, a single modality—visual, textual, or acoustic—can dominate; in others, two modalities jointly align with the overall sentiment label. For example, in the third sample in Fig. 1, visual and textual modalities are both dominant. When modal dominance is not fixed, tri-modality models adapt poorly, and text-centric models may be distracted by secondary information, introducing bias. Current methods lack mechanisms to identify and adjust for shifting modal dominance and, in the regression stage, largely rely on symmetric loss functions, leaving asymmetric prediction risks unaddressed.

To address the above challenges, we introduce the Three-Way Decision (**3WD**) Enhanced **D**ynamic **R**outing **T**ransformer (**3WD-DRT**) framework, which adopts Three-Way Decision theory [4] as a foundational principle to develop a quality-aware and decision-guided end-to-end learning method. This theory partitions the decision space into acceptance, rejection, and deferment regions, offering a principled methodology for managing uncertainty in complex decision-making processes. Building on this, our framework

incorporates the decision-making mechanism throughout three critical stages of its architecture: first, it enhances the quality of input representations via a pretraining task guided by knowledge alignment supervision; second, it adaptively integrates multimodal information through a dynamic routing fusion strategy; and finally, it conducts emotion regression with improved precision and robustness using a cost-sensitive loss function tailored to the task's specific demands. Our main contributions can be summarized at four levels:

i. Stage I (Unimodal Knowledge Enhancement): we design a dual self-supervised pre-training task incorporating a Knowledge Alignment Loss to ensure high-quality and highly discriminative input features through a three-way decision-based quality assessment of injected domain knowledge.
ii. Stage II (Three-Way Decision-Based Dynamic Fusion): we propose a dynamic routing fusion mechanism that, based on the confidence of each modality, utilizes a three-way decision gate to adaptively amplify, regulate, or suppress its contribution, thereby achieving robust and efficient feature fusion.
iii. Stage III (Cost-Sensitive Sentiment Regression): we design a novel loss function, the cost-sensitive loss: $\mathcal{L}_{3WD}$, which introduces learnable decision boundaries to partition the semantic region of predicted values into three sections and imposes asymmetric prediction penalties, making the model more sensitive to severe polarity misclassifications.
iv. Extensive experiments were conducted on four multimodal sentiment analysis benchmarks, CH-SIMS, CH-SIMSv2, MOSI, and MOSEI, which consistently demonstrated the superiority of the proposed 3WD-DRT framework. On CH-SIMS, it achieves the highest Pearson correlation (0.627), Acc-2 (81.37 %), and F1-score (81.91 %), while matching the best MAE (0.408). For the more challenging CH-SIMSv2, 3WD-DRT sets new state-of-the-art results, including an F1-score of 84.01 %, Corr of 0.764, and Acc-2 of 83.05 %. On English-language datasets, the framework attains an F1-score of 87.08 % on MOSI and 88.26 % on MOSEI.

The remainder of this paper is organized as follows. Section 2 reviews related work on the theoretical foundations of three-way decisions and recent advances in multimodal fusion, representation learning, uncertainty modeling, and cost-sensitive learning. Section 3 presents the proposed methodology, including the problem formulation, overall architecture with three progressive stages (unimodal knowledge enhancement, three-way decision-based dynamic fusion, and cost-sensitive sentiment regression), and the joint optimization and training algorithms. Section 4 reports extensive experiments, covering datasets, baselines, main results, significance testing, ablation and sensitivity analyses, efficiency and robustness studies, and a case study. Section 5 concludes the paper and discusses future directions.

## 2. Related work

This section establishes the context for our framework through a structured review organized into three parts. First, we delve into the theoretical foundations and developments of 3WD theory. We then examine how these principles have been applied in unimodal sentiment analysis to manage uncertainty. Finally, the chapter broadens its scope to survey the diverse fusion techniques currently employed in MSA, analyzing their inherent limitations. By identifying the critical research gaps that emerge from the intersection of these domains, this three-part review systematically motivates the novel contributions of our study.

### 2.1. The development of the three-way decision theoretical framework

Three-way decision (3WD), originating from rough set theory [4], provides a principled framework for managing uncertainty by partitioning the universal set into acceptance (ACC), rejection (REJ), and deferment (DEF) regions. Professor Yao laid the theoretical foundation of 3WD [4] and, from a three-valued philosophical perspective, connected it to explainable artificial intelligence, thereby advancing the transparency of AI systems [5]. To accommodate more complex decision-making, researchers further extended 3WD to multisource information systems, such as interval-valued models [6] and fuzzy set integrations [7], which enhanced its ability to address boundary uncertainty.

Building on these advances, 3WD has been increasingly adopted in machine learning for tasks such as information filtering [8]. Although not directly designed for sentiment analysis, these studies demonstrate 3WD's versatility in integrating diverse sources and managing uncertainty, offering methodological insights for multimodal sentiment research.

### 2.2. Applications of three-way decisions in unimodal sentiment analysis

Given the inherent ambiguity of sentiment expressions (e.g., neutral or contradictory statements), 3WD provides a natural paradigm for unimodal sentiment analysis. By introducing a deferment option, it accommodates uncertain samples that challenge binary classifiers. Chen et al.'s seminal three-way sentiment classification model [9] illustrates this principle, improving robustness through boundary–region processing. Building on this foundation, further studies refined decision mechanisms with Decision-Theoretic Rough Sets, yielding reliable classification rules and threshold mappings for context-dependent uncertainty [10,11].

Beyond sentiment analysis, theoretical innovations such as multi-granularity analysis and hierarchical metrics continue to enrich 3WD. Meanwhile, recent integrations with deep learning, such as Wang et al.'s DeeBERT-S3WD model [12] and knowledge-enhanced sequential decision frameworks [13], highlight the adaptability of 3WD to modern AI paradigms.

These developments establish the theoretical value and practical utility of 3WD for modeling sentiment uncertainty. Nevertheless, applications have remained predominantly unimodal, and its potential in multimodal scenarios—where richer information, intricate

interactions, and diverse uncertainty sources coexist—remains underexplored. Addressing this gap is a central motivation of our study.

### 2.3. Fusion techniques in multimodal sentiment analysis

A central challenge in MSA lies in the effective integration of heterogeneous modalities. Early fusion methods such as Tensor Fusion Networks (TFN) and Low-rank Multimodal Fusion (LMF) established the foundation by modeling interactions across modalities through explicit tensor-based or factorized operations [19]. These approaches demonstrated the value of joint feature learning but were limited by their computational complexity and relatively shallow fusion depth.

To overcome these constraints, Transformer-based models have brought multimodal fusion into a new stage. MulT [2] and its derivatives exemplify the capacity of cross-modal attention mechanisms to capture fine-grained temporal and semantic dependencies. These advances have significantly enriched the representational capacity of MSA.

However, most existing techniques are inherently static, embedding fixed hierarchical assumptions regarding modality interactions. For example, concatenation-based fusion [20] and text-centered attention schemes [3] often presuppose the dominance of text, which can bias decision-making in contexts where visual or acoustic cues are more informative, such as sarcasm or irony. Even knowledge-guided fusion methods [18], though adaptive, are constrained by their dependence on external resources that may introduce noise or bias. These limitations underscore the broader challenge of handling modality dynamics in a principled manner.

To address these issues, researchers have increasingly turned to dynamic fusion mechanisms. Dynamic routing has emerged as a promising direction, with implementations such as Capsule Networks [21], the Modality Focusing Model [22], and attention-based routing designs [23], all aiming to adapt computational flows to input-specific characteristics. Despite these innovations, current solutions often lack a solid theoretical foundation for balancing flexibility with interpretability.

Recent surveys further contextualize these developments. Pandey et al. [24] mapped its technological evolution. More recent work has highlighted opportunities arising from large-scale vision-language pretrained models [25] and diverse deep architectures [26]. Yet, persistent challenges remain, including cross-modal ambiguity and over-reliance on textual modalities.

The above insights underscore the pressing need for a theoretically grounded fusion paradigm that can adaptively accommodate shifting modality contributions. This challenge directly motivates the design of our proposed 3WD-DRT framework.

### 2.4. Representation learning and uncertainty modeling

Beyond fusion strategies, two parallel lines of research—representation learning and uncertainty modeling—have become central to advancing MSA. Representation learning seeks to improve the quality of input features, while uncertainty modeling equips models with the ability to assess their predictive confidence.

In representation learning, significant progress has been made in reducing cross-modal redundancy and enhancing feature expressiveness. The Modality-Invariant and -Specific representation framework (MISA) [16] disentangles features into invariant and modality-specific components, yielding more distinguishable signals for fusion. Self-supervised methods further enhance unimodal representations, as in Self-MM [17], which leverages pseudo-labels to capture discriminative modality-specific characteristics. Similarly, the Multimodal Mutual Information Maximization model (MMIM) [27] preserves task-relevant features by maximizing mutual information between unimodal and fused representations. Collectively, these approaches provide richer and more structured features, yet downstream decision modules typically remain risk-neutral, overlooking potential uncertainty in combining signals across modalities.

In parallel, uncertainty modeling has emerged as an effective tool for handling modality conflict and decision risk. Evidential Deep Learning (EDL) [28] and Bayesian Deep Learning [29] offer principled means of quantifying epistemic and aleatoric uncertainty. Building on these foundations, the Uncertainty-aware Late Fusion model with Hybrid Uncertainty Calibration (ULF-HUC) [15] applies EDL-based estimation to each unimodal branch, prioritizing modalities with lower uncertainty when predictions diverge. While this strategy introduces principled uncertainty reasoning into multimodal arbitration, its operation is largely post hoc: uncertainty estimation is confined to independent unimodal outputs and affects only the final decision stage. Consequently, it does not guide the deeper processes of representation learning or fusion.

Representation learning and uncertainty modeling have each substantially advanced multimodal sentiment analysis: the former by refining feature quality, the latter by calibrating predictive confidence. Yet these two trajectories have largely evolved in parallel. What remains absent is a unified framework that embeds uncertainty estimation into the very processes of feature learning and dynamic fusion, rather than treating it as a peripheral safeguard.

### 2.5. Cost-sensitive learning

A central premise in conventional machine learning is that all misclassification errors are equally costly. This assumption rarely holds in practice, where class imbalance often skews predictions toward the majority class, yielding deceptively high accuracy [30]. To mitigate such limitations, Cost-Sensitive Learning (CSL) was introduced to incorporate asymmetric penalties, shifting the optimization focus from minimizing error rate to minimizing the expected total cost [31]. CSL methods are broadly categorized into direct approaches, which are intrinsically cost-sensitive, and wrapper approaches, which adapt cost-insensitive models to account for varying error costs [30].

Over time, CSL has matured into a rich field addressing diverse challenges. With the rise of deep learning, CSL has been integrated into neural architectures, for instance, through cost-sensitive residual networks with weighted loss layers [32]. CSL has further been

extended to areas like active learning [33] and advanced hierarchical decision systems [34]. Collectively, these studies highlight the versatility of CSL, with the majority of applications concentrated on classification problems involving imbalanced data.

By contrast, its potential in multimodal affective tasks remains largely unexplored. In particular, sentiment regression entails a unique form of semantic asymmetry: predicting a strongly positive sample as strongly negative is far more costly than small deviations in intensity. Yet, few studies have applied CSL principles to explicitly design loss functions that capture such asymmetric risks in multimodal affective regression. Addressing this gap constitutes one of the key contributions of our work.

A detailed analysis of the current research landscape reveals that, despite significant advances in both MSA and 3WD theory, their integration remains insufficiently explored. Three key gaps can be identified:

i. **Theoretical and architectural disconnection:** Most existing applications of 3WD are confined to post-hoc decision-level adjustments in unimodal classification tasks [9]. The core principle of 3WD partitioning decisions into acceptance, rejection, or deferment based on uncertainty—has rarely been embedded within the internal computation processes of neural models. Present MSA systems generally lack mechanisms for dynamically routing modality information based on 3WD, and few leverage it to guide representation learning or assess feature quality.
ii. **Structural shortcomings in fusion mechanisms:** While several dynamic fusion approaches have emerged [18,22], many depend heavily on external knowledge or lack a robust, theory-driven arbitration process. Current models typically do not include a principled gating module that evaluates modality confidence from input data and modulates information flow accordingly—something that 3WD could systematically support.
iii. **Semantic insensitivity of loss functions:** In sentiment regression, loss functions such as Mean Squared Error (MSE) remain standard. These symmetric functions treat all prediction errors uniformly and fail to reflect the inherent asymmetry of sentiment tasks—for instance, predicting "positive" as "negative" is far more consequential than a minor deviation in intensity. Existing loss functions overlook such semantic disparities.

To address these gaps, we introduce the 3WD-DRT, a unified framework designed to systematically tackle the architectural, fusion, and regression challenges in multimodal sentiment analysis. For the issues of architectural inconsistency and fusion inadequacy, we propose a dynamic fusion mechanism that integrates 3WD-based gating. This mechanism adaptively regulates information flow by leveraging learned confidence scores from each modality. Complementing this, we implement a dual self-supervised pretraining strategy that incorporates a knowledge-aligned loss ($\mathcal{L}_{ka}$), offering enhanced feature representations and tighter alignment with the 3WD framework. To address the semantic limitations in conventional regression loss functions, we explore the underdeveloped dimension of cost-sensitive regression. Specifically, we develop a novel loss function, $\mathcal{L}_{3WD}$, which segments the continuous sentiment space into semantically meaningful regions using learnable decision boundaries. This enables a nonlinear penalty structure that assigns greater costs to critical errors, such as polarity inversions, thereby improving both model robustness and predictive accuracy.

As shown in Table 1, 3WD-DRT presents a sharp contrast to current mainstream SOTA models in terms of fusion mechanism, decision processing, and modality confidence evaluation, offering a more comprehensive and robust solution.

**Table 1**
Comparison of representative models in fusion strategy, modality confidence handling, and decision processing.

| Model | Core idea | Fusion type | Modality confidence handling | Decision processing |
|---|---|---|---|---|
| MulT [2] | Deep cross-modal fusion based on Transformer | Tensor attention-based fusion | Implicitly encoded via attention weights; lacks explicit confidence estimation | Standard regression loss (MSE) |
| ALMT [14] | Language-guided suppression of cross-modal noise | Transformer fusion guided by text modality | Fixed text-centric hierarchical structure | Standard regression loss |
| ULF-HUC [15] | Late fusion based on evidential theory | Rule-based late fusion at the decision layer | Modality-specific uncertainty scores derived from evidential theory | Conflict arbitration at the decision level based on uncertainty |
| MISA [16] | Disentanglement of modality-invariant and -specific representations | Concatenation of disentangled features | Not applicable (modalities treated equally post-disentanglement) | Standard regression loss |
| Self-MM [17] | Self-supervised learning to enhance modality representation | Linear fusion after feature concatenation | Implicitly reflected through auxiliary task loss weights | Standard regression loss |
| KuDA [18] | Knowledge-guided dynamic attention mechanism | Dynamic attention fusion | Dynamic, guided by external affective knowledge base | Standard regression loss |
| **3WD-DRT (Ours)** | Systematic integration of three-way decision theory | **3WD-gated dynamic routing** | Dynamic, driven by data-intrinsic modality confidence prediction | Dynamic routing (fusion) and cost-sensitive loss (decision) |

In summary, the 3WD-DRT framework operationalizes 3WD theory across representation learning, information fusion, and regression stages, offering an integrated and theoretically informed solution to the core limitations in current multimodal sentiment analysis research.

## 3. Method

The methodology of our proposed 3WD-DRT framework is designed to be comprehensive, integrating several novel components across a multi-stage architecture. To provide a clear roadmap for the reader, this section is structured to follow the flow of information as it is processed by our model, from initial unimodal encoding to the final cost-sensitive prediction.

Fig. 2 serves as the central schematic overview for this entire section. It illustrates the three core stages of our framework, which will be detailed sequentially in the following subsections:

Stage 1: Unimodal Knowledge Enhancement (Section 3.3.1): We first describe how each modality's representation is independently enriched with domain-specific and domain-general knowledge during a self-supervised pre-training phase.

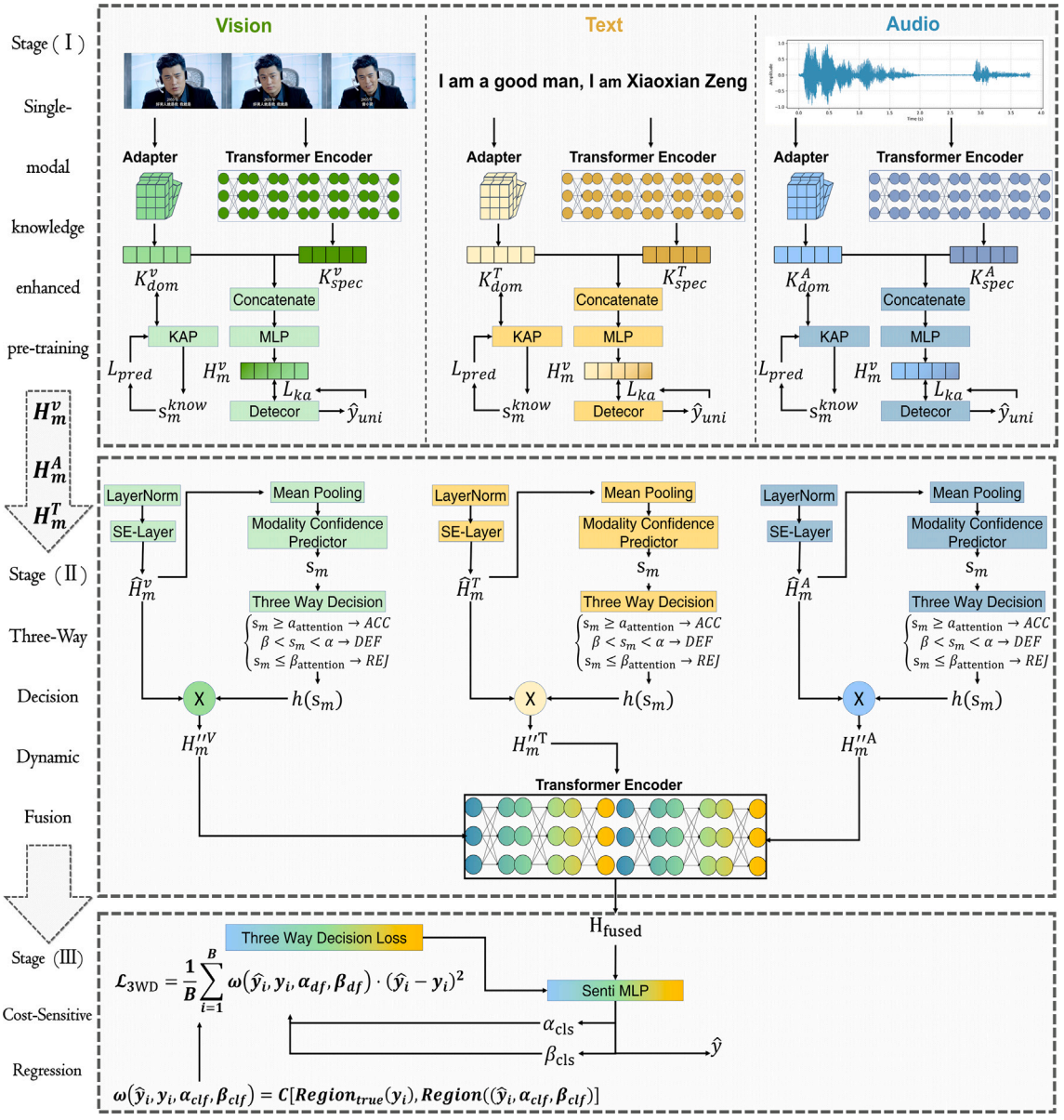


**Fig. 2.** The 3WD-DRT framework: (a) Stage 1 ensures representation quality via dual self-supervision ($\mathcal{L}_{KA}$); (b) Stage 2 weights modalities by confidence $s_m$ for Transformer fusion; (c) Stage 3 achieves risk-aware regression using $\mathcal{L}_{3WD}$ with boundaries $\alpha_{clf}, \beta_{clf}$.

**Table 2**
Symbol Descriptions and Corresponding Modules.

| Symbol | Description | Relevant Module/Stage |
|---|---|---|
| $D$, $X_i$, $y_i$ | Dataset, the $i$-th multimodal sample, ground-truth label | Problem Formulation |
| $m$, $X_i^m$ | Modality index ($T$, $V$, $A$), $m$-th modality features of the $i$-th sample | Problem Formulation |
| $\hat{y}_i$, $\Theta$, $F$ | Predicted sentiment value, model parameter set, overall mapping function | Problem Formulation, Stage III |
| *Stage I: Uni-modal Knowledge-Augmented Pretraining* | | |
| $K_{\text{spec}}$, $K_{\text{dom}}$ | Modality-specific knowledge, domain-general knowledge | Uni-modal Knowledge Augmentation Path |
| $H'_m$ | Enhanced unimodal feature representation | Uni-modal Knowledge Augmentation Path |
| $s_m^{\text{know}}$ | Knowledge alignment confidence score | Knowledge Alignment Predictor (KAP) |
| $\alpha$, $\beta$ | Decision thresholds for knowledge alignment loss | Knowledge Alignment Loss |
| $\hat{y}_{\text{uni}}$ | Predicted sentiment in the unimodal pretraining phase | Self-Supervised Predictor (Decoder) |
| *Stage II: Three-Way Decision-Based Dynamic Fusion* | | |
| $\hat{H}_m$ | Preprocessed unimodal features (LN, SE) | Uni-modal Dynamic Gating |
| $s_m$ | Modality confidence score | Modality Confidence Predictor (MCP) |
| $\alpha_{\text{attn}}$, $\beta_{\text{attn}}$ | Learnable decision thresholds for fusion gating | Three-Way Decision Logic |
| $\gamma_{\text{ACC/DEF/REJ}}$ | Learnable scaling coefficients for decision regions | Dynamic Re-weighting |
| $h(s_m)$ | Dynamic scaling factor | Dynamic Re-weighting |
| $H''_m$ | Re-weighted unimodal features | Dynamic Re-weighting |
| $H_{\text{fused}}$ | Final fused multimodal representation | Multimodal Fusion |
| *Stage III: Cost-Sensitive Regression* | | |
| $\alpha_{\text{clf}}$, $\beta_{\text{clf}}$ | Learnable decision thresholds for regression | Regression Head (SentiMLP) |
| Region$(\cdot)$ | Region classification function | Cost-Sensitive Loss |
| $C$ | Predefined cost matrix | Cost-Sensitive Loss |
| $w(\cdot)$ | Cost-sensitive weight retrieved from $C$ | Cost-Sensitive Loss |
| *Loss Functions* | | |
| $\mathcal{L}_{\text{KA}}$, $\mathcal{L}_{\text{pred}}$ | Knowledge alignment loss, reconstruction prediction loss | Stage I |
| $\mathcal{L}_{\text{TD-attn}}$ | Threshold regularization loss | Stage II |
| $\mathcal{L}_{\text{3WD}}$ | Cost-sensitive task loss | Stage III |
| $\mathcal{L}_{\text{total}}$ | Final total loss for Stage II fine-tuning | Final Optimization Objective |
| $\lambda(\cdot)$ | Weighting hyperparameters for loss components | Loss Functions Across Stages |

Stage 2: Three-Way Decision-Based Dynamic Fusion (Section 3.3.2): Next, we explain the core mechanism of our framework, detailing how modality confidence is estimated and used by a three-way decision gate to dynamically re-weight and fuse the multimodal features.

Stage 3: Cost-Sensitive Sentiment Regression (Section 3.3.3): Finally, we present our novel cost-sensitive loss function, explaining how it evaluates the fused representation to produce a final prediction that is sensitive to the asymmetric risks of semantic errors.

We begin with a formal problem formulation and a list of notations to establish the necessary groundwork.

### 3.1. Problem formulation

This study addresses multimodal sentiment regression. Given a dataset $\mathcal{D} = \{(X_i, y_i)\}_{i=1}^N$ with $N$ samples, each sample $X_i$ contains feature sequences from three modalities: $X_i = \{X_i^T, X_i^V, X_i^A\}$ for text, visual, and audio inputs.

For modality $m \in \{T, V, A\}$, input features form a 2D tensor $X_i^m \in \mathbb{R}^{L_m \times d_m}$, where $L_m$ is sequence length and $d_m$ is feature dimension. Each $X_i$ has a continuous sentiment label $y_i \in \mathbb{R}$, typically normalized (e.g., to $[-1, 1]$).

We learn a nonlinear mapping $\mathcal{F}$ parameterized by $\Theta$ that predicts sentiment scores:

$$\hat{y}_i = \mathcal{F}(X_i; \Theta) \tag{1}$$

Our objective minimizes a composite loss $\mathcal{L}_{\text{total}}$ based on Three-Way Decision theory, which handles asymmetric prediction risks.

### 3.2. List of notations

To facilitate the reader's understanding of the methodology, we summarize the primary mathematical notation used throughout this section. Detailed definitions are provided in Table 2.

To further clarify the roles of the various thresholds used throughout our framework, it is important to note that the three pairs of $\alpha$ and $\beta$ thresholds listed in the Table 2 are distinct, operate on different inputs, and are governed by different mechanisms. Specifically:

- The pre-training thresholds ($\alpha$ and $\beta$) are hyperparameters used to assess knowledge quality in the $\mathcal{L}_{KA}$ loss.
- The fusion gating thresholds ($\alpha_{\text{attn}}$, $\beta_{\text{attn}}$) are independently learned parameters that operate on modality confidence scores $s_m$ to dynamically re-weight features.
- The regression loss thresholds ($\alpha_{\text{clf}}$, $\beta_{\text{clf}}$) are also independently learned parameters, but they operate on the final sentiment scores $(\hat{y}, y)$ to apply asymmetric costs in the $\mathcal{L}_{\text{3WD}}$ loss function.

### 3.3. Overall architecture and design principles

To handle modality noise, heterogeneity, and asymmetric risks, we propose the **Three**-**W**ay **D**ecision Enhanced **D**ynamic **R**outing **T**ransformer (**3WD-DRT**) framework. As illustrated in Fig. 2. Its "quality-aware, decision-driven" paradigm integrates 3WD throughout the entire processing pipeline, enabling progressive information filtering, dynamic modality weighting, and risk-aware prediction. The architecture has three stages: unimodal knowledge enhancement, Three-Way Decision-Based Dynamic Fusion, and cost-sensitive regression, trained via a decoupled three-stage strategy.

This design is particularly crucial for the stability and effectiveness of 3WD-DRT. At the core of our architecture lies the Modality Confidence Predictor (MCP), which estimates the reliability of each modality for dynamic fusion. If trained end-to-end from scratch, the MCP would simultaneously attempt to learn weighting functions while the unimodal encoders are still producing noisy, undertrained features. This co-adaptation easily destabilizes optimization, often leading to unreliable confidence scores and poor convergence. By decoupling training, Stage 1 ensures that unimodal encoders become competent predictors independently, producing semantically aligned and high-quality representations. When entering Stage 2, the MCP and fusion modules are then provided with a stable and meaningful feature space. Their task is reduced to the more focused problem of distinguishing and re-weighting already strong signals rather than compensating for encoder immaturity. This staged training stabilizes confidence prediction, mitigates gradient noise, and allows the fusion mechanism to converge more effectively.

#### 3.3.1. Stage 1: unimodal knowledge enhancement

During pretraining, we develop domain-generalizable representations for each modality. A backbone encoder processes input $X_m$ to extract two knowledge types: *specific knowledge* ($K_{\text{spec}}$) from top layers containing contextual semantics, and *domain-generalizable knowledge* ($K_{\text{dom}}$) obtained via a `Adapter`. This adapter employs a bottleneck structure (dimension reduction → self-attention → expansion) to distill cross-layer features.

We concatenate and project $[K_{\text{spec}}; K_{\text{dom}}]$ through an MLP to generate enhanced representations $H'_m$. Training combines two objectives: First, a Knowledge Alignment Predictor scores quality $s_m^{\text{know}}$ for $K_{\text{dom}}$ supervised by:

$$\mathcal{L}_{KA} = \begin{cases} w_{\text{ACC}} \cdot \text{MSE}(s_m^{\text{know}}, 1) & s_m^{\text{know}} \geq \alpha \\ w_{\text{DEF}} \cdot \text{MSE}(s_m^{\text{know}}, 1) & \beta < s_m^{\text{know}} < \alpha \\ w_{\text{REJ}} \cdot \text{MSE}(s_m^{\text{know}}, 1) & s_m^{\text{know}} \leq \beta \end{cases} \tag{2}$$

with $w_{\text{REJ}} > w_{\text{DEF}} > w_{\text{ACC}}$ penalizing low-quality knowledge. Specifically, we set $w_{\text{REJ}} = 5.0$, $w_{\text{DEF}} = 2.0$, and $w_{\text{ACC}} = 1.0$, following a principled risk hierarchy inspired by Three-Way Decision theory. Rejection receives the highest penalty as it corresponds to severely misaligned knowledge, deferment incurs a moderate penalty for uncertain cases, and acceptance carries the lowest penalty for well-aligned knowledge. These fixed ratios serve as *structural constants* rather than tunable hyperparameters, since pretraining functions as a proxy task whose primary role is to provide the encoder with a stable, consistent, and challenging learning signal. Empirically, the ratio 5:2:1 balances optimization stability and effective risk separation, whereas steeper (e.g., 10:3:1) or narrower (e.g., 2:1.5:1) settings destabilize convergence or weaken region distinctions. Importantly, because the purpose of these weights is to shape representation learning during pretraining, their exact values have only marginal influence on downstream performance compared to parameters actively optimized in fine-tuning.

Second, a decoder reconstructs sentiment labels from $H'_m$ using standard MSE loss $\mathcal{L}_{\text{pred}}$:

$$\mathcal{L}_{\text{pred}} = \frac{1}{n} \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2. \tag{3}$$

The joint loss $\mathcal{L}_{\text{stage1}} = \lambda_{KA}\mathcal{L}_{KA} + \lambda_{\text{pred}}\mathcal{L}_{\text{pred}}$ guides optimization, with best encoders preserved for Stage 2.

#### 3.3.2. Stage 2: three-way decision-based dynamic fusion

In this fine-tuning stage, we integrate dynamic fusion with cost-sensitive regression. At the core is a three-way decision gate that adjusts the contribution of each modality according to its estimated reliability. The procedure comprises four steps:

i. Modality Confidence Estimation: For the feature representation of each modality $H_m$, Layer Normalization and an SE-Layer are applied to produce an enhanced form $\hat{H}_m$. A Modality Confidence Predictor (MCP)—a lightweight MLP—takes the mean-pooled features $\overline{\hat{H}}_m$ and outputs a scalar confidence score $s_m \in [0, 1]$. This score reflects the model's estimation of a modality's quality and relevance for a given sample.

ii. Three-Way Decision Gating and Dynamic Re-weighting: Using the confidence score $s_m$, each modality is routed into one of three categories: Acceptance (high confidence), Deferment (medium confidence), or Rejection (low confidence). The boundaries are set by two learnable thresholds, $\alpha_{\text{attn}}$ and $\beta_{\text{attn}}$. Rather than discarding data via hard gating, we employ a dynamic re-weighting function $h(s_m)$ to scale each modality's contribution smoothly:

Acceptance ($s_m \geq \alpha_{\text{attn}}$): Features are amplified to emphasize their influence. Deferment ($\beta_{\text{attn}} < s_m < \alpha_{\text{attn}}$): Features are scaled moderately, preserving potentially useful signals while limiting noise. Rejection ($s_m \leq \beta_{\text{attn}}$): Features are heavily attenuated due to low reliability.

The scaling process is defined as

$$h(s_m) = \begin{cases} \gamma_{\text{ACC}} \cdot s_m & \text{if } s_m \geq \alpha_{\text{attn}}, \\ \gamma_{\text{DEF}} \cdot s_m & \text{if } \beta_{\text{attn}} < s_m < \alpha_{\text{attn}}, \\ \gamma_{\text{REJ}} \cdot s_m & \text{if } s_m \leq \beta_{\text{attn}}. \end{cases} \tag{4}$$

where $\gamma_{\text{ACC}}$, $\gamma_{\text{DEF}}$, and $\gamma_{\text{REJ}}$ are learnable parameters that control the scaling strength for each region. The re-weighted features are given by $H''_m = \hat{H}_m \cdot h(s_m)$.

iii. Learnable Thresholds and Stabilization: The thresholds $\alpha_{\text{attn}}$ and $\beta_{\text{attn}}$ are learnable parameters, initialized to 0.7 and 0.3. They are optimized jointly with all model parameters via backpropagation using the overall task loss $\mathcal{L}_{\text{total}}$.

To maintain a valid deferment range ($\alpha_{\text{attn}} > \beta_{\text{attn}}$), we introduce a threshold regularization loss $\mathcal{L}_{\text{TD-attn}}$:

$$\mathcal{L}_{\text{TD-attn}} = \text{LeakyReLU}(\beta_{\text{attn}} - \alpha_{\text{attn}} + \delta) \tag{5}$$

where $\delta$ is a small margin (e.g., 0.05). This term discourages threshold collapse or inversion during training.

iv. Final Fusion: The adjusted features $H''_T, H''_V, H''A$ are pooled, concatenated, and processed by a final TransformerEncoder to obtain the fused multimodal representation $H$fused, which is then passed to the regression head.

Weighted features $\{H''_T, H''_V, H''_A\}$ undergo pooling, concatenation, and `TransformerEncoder` processing to generate $H_{\text{fused}}$. For regression, the MLP predicts sentiment score $\hat{y}$ and thresholds $\alpha_{clf}$, $\beta_{clf}$ with Cost-Sensitive Loss.

### 3.3.3. Stage 3: cost-sensitive sentiment regression

i. **Regression**: $H_{\text{fused}}$ enters the `SentiMLP` regression head (MLP), outputting: (a) predicted sentiment $\hat{y}$, and (b) learnable thresholds $\alpha_{clf}, \beta_{clf}$ partitioning the prediction space.

ii. **Cost-Sensitive Loss:** Our $\mathcal{L}_{3WD}$ uses a region function:

$$\text{Region}(v, \alpha, \beta) = \begin{cases} \text{ACC} & v \geq \alpha \\ \text{REJ} & v \leq \beta \\ \text{DEF} & \text{otherwise} \end{cases} \tag{6}$$

This definition explicitly assigns predictions that fall exactly on a boundary to a non-deferred region: a value equal to $\alpha_{clf}$ is considered an Acceptance (ACC), and a value equal to $\beta_{clf}$ is considered a Rejection (REJ).

Task loss $\mathcal{L}_{\text{3WD}}$ is a weighted MSE:

$$\mathcal{L}_{\text{3WD}} = \frac{1}{B} \sum_{i=1}^{B} w(\hat{y}_i, y_i, \alpha_{clf}, \beta_{clf}) \cdot (\hat{y}_i - y_i)^2 \tag{7}$$

where weight $w$ depends on a cost matrix $C$ comparing true and predicted regions:

$$w = C\left[\text{Region}(y_i, \alpha_{clf}, \beta_{clf}), \text{Region}(\hat{y}_i, \alpha_{clf}, \beta_{clf})\right] \tag{8}$$

## 3.4. Joint optimization objective

The joint fine-tuning objective combines:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{3WD}} + \lambda_{td} \cdot \mathcal{L}_{TD-attn} + \lambda_{nce} \cdot \mathcal{L}_{nce} \tag{9}$$

where $\mathcal{L}_{\text{3WD}}$ is the core regression loss, $\mathcal{L}_{TD-attn}$ stabilizes fusion thresholds, and $\mathcal{L}_{nce}$ (optional) is a standard contrastive loss [35]:

$$\mathcal{L}_{\text{nce}} = -\log \frac{\exp(\sin(z_i, z_p)/\tau)}{\sum_{k=1}^{K} \exp(\sin(z_i, z_k)/\tau)}. \tag{10}$$

## 3.5. Training algorithms

To ensure the reproducibility of the proposed method and to clearly articulate the model's training paradigm, we provide detailed algorithmic procedures in the form of pseudocode. As previously described, our 3WD-DRT framework adopts a decoupled three-stage training strategy.

Algorithm 1 depicts Stage 1, where unimodal knowledge-enhancement pretraining is conducted. In this stage, each modality is optimized independently. For each modality—text ($T$), vision ($V$), and audio ($A$)—we construct a dedicated knowledge-enhanced subnetwork to perform self-supervised learning. Upon completion, the optimal encoder weights for each modality are retained and used to initialize the multimodal model in the next stage.

Algorithm 2 presents Stage 2 and Stage 3, the end-to-end multimodal fine-tuning process. In this stage, all components are jointly optimized to accomplish the final regression task. The model $M_{3WD\text{-}DRT}$ is initialized with the pretrained weights and trained on multimodal inputs $(X^T, X^V, X^A)$. Each modality is first processed by its corresponding encoder UniEncKI to obtain the enhanced intermediate features $H'T, H'V, H'A$. These representations are then fused via the dynamic multimodal fusion module DyMultiFus,

**Algorithm 1** Stage 1 - Unimodal knowledge enhancement pretraining.

**Require:** Data loader `Dataloader`
**Ensure:** Optimal unimodal encoder weights $\Theta^*_{enc\_m}$
**for** modality $m$ in $\{T, V, A\}$ **do**
  Initialize $M_{\text{pretrain\_m}} \leftarrow \text{UniPretrain}(m)$
  Initialize $\text{Optimizer}_m$
  **for** epoch = 1 to `N_epochs_pretrain` **do**
    **for** batch $(X^m, y)$ in `Dataloader` **do**
      $\hat{y}_{\text{uni}}, s_m^{\text{know}} \leftarrow M_{\text{pretrain\_m}}(X^m)$
      $\mathcal{L}_{\text{pred}} \leftarrow \text{MSE}(\hat{y}_{\text{uni}}, y)$
      $\mathcal{L}_{KA} \leftarrow \text{KnowledgeAlignmentLoss}(s_m^{\text{know}})$
      $\mathcal{L}_{\text{stage1}} \leftarrow \mathcal{L}_{\text{pred}} + w_{ka} \cdot \mathcal{L}_{KA}$
      Backpropagate and update $\text{Optimizer}_m$
    **end for**
  **end for**
  Save best encoder $\Theta^*_{enc\_m}$
**end for**
**return** $\{\Theta^*_{enc\_T}, \Theta^*_{enc\_V}, \Theta^*_{enc\_A}\}$

**Algorithm 2** Stage 2 and 3: dynamic fusion and regression.

**Require:** Data loader `Dataloader`, pretrained weights $\{\Theta^*_{enc\_m}\}$
**Ensure:** Trained model $M_{3WD-DRT}$
Initialize $M_{3WD-DRT}$
Load weights: $M_{3WD-DRT}$.load_state_dict($\{\Theta^*_{enc\_m}\}$)
Initialize `Optimizer`
**for** epoch = 1 to `N_epochs_finetune` **do**
  **for** batch $(\{X^T, X^V, X^A\}, y)$ in `Dataloader` **do**
    ▷ 1. Unimodal encoding
    $\{H'_T, H'_V, H'_A\} \leftarrow M_{3WD-DRT}$.UniEncKI($\{X^T, X^V, X^A\}$)
    ▷ 2. Dynamic fusion
    $H_{\text{fused}}, \mathcal{L}_{TD-attn}, \mathcal{L}_{nce} \leftarrow M_{3WD-DRT}$.DyMultiFus($\{H'_T, H'_V, H'_A\}$)
    ▷ 3. Regression
    $\hat{y}, \alpha_{clf}, \beta_{clf} \leftarrow M_{3WD-DRT}$.SentiMLP($H_{\text{fused}}$)
    ▷ 4. Loss calculation
    $\mathcal{L}_{\text{3WD}} \leftarrow \mathcal{L}_{3WD}(\hat{y}, y, \alpha_{clf}, \beta_{clf})$
    $\mathcal{L}_{\text{total}} \leftarrow \mathcal{L}_{\text{3WD}} + \lambda_{td} \cdot \mathcal{L}_{TD-attn} + \lambda_{nce} \cdot \mathcal{L}_{nce}$
    ▷ 5. Optimization
    Backpropagate and update `Optimizer`
  **end for**
**end for**
**return** trained model $M_{3WD-DRT}$

which is guided by the three-way decision mechanism to produce the final fused representation $H\text{fused}$. Simultaneously, two auxiliary losses are computed: the decision-attention loss $\mathcal{L}_{TD-attn}$ and the contrastive learning loss $\mathcal{L}_{nce}$. The fused representation is passed to the classification module SentiMLP to generate the regression output $\hat{y}$ along with the modality-adaptive decision thresholds $\alpha_{clf}$ and $\beta_{clf}$. Based on these outputs, the primary loss $\mathcal{L}_{3WD}$ is computed, formulated as a three-way-decision-aware regression loss. Throughout the training process, the model is optimized via this composite loss function until convergence on the validation set, resulting in the final multimodal sentiment regression model $M_{3WD\text{-}DRT}$.

## 4. Experiments

### 4.1. Datasets and evaluation metrics

We evaluate our method on four widely used MSA benchmark datasets: CH-SIMS (Yu et al. [1]), CH-SIMSv2 (Liu et al. [36]), MOSI (Zadeh et al. [37]), and MOSEI (Zadeh et al. [38]). Table 3 summarizes their statistics.

While both CH-SIMS and CH-SIMSv2 are Chinese MSA benchmarks derived from movie/TV clips, they exhibit key differences relevant to model evaluation:

**Table 3**
The statistics of four datasets.

| Dataset | Train | Valid | Test | Total | Language |
|---|---|---|---|---|---|
| CH-SIMS | 1368 | 456 | 457 | 2281 | Chinese |
| CH-SIMSv2 | 2722 | 647 | 1034 | 4403 | Chinese |
| MOSI | 1284 | 229 | 686 | 2199 | English |
| MOSEI | 16,326 | 1871 | 4659 | 22,856 | English |

**Table 4**
Hyperparameter configurations used for experiments on four benchmark datasets.

| Hyperparameter | CH-SIMS | CH-SIMSv2 | MOSI | MOSEI |
|---|---|---|---|---|
| *Training & Model Parameters* | | | | |
| Batch Size | 32 | 32 | 32 | 64 |
| Initial Learning Rate | $3 \times 10^{-5}$ | $3 \times 10^{-5}$ | $3 \times 10^{-5}$ | $4 \times 10^{-5}$ |
| Fusion Transformer Layers | 3 | 3 | 2 | 4 |
| Hidden Dimension ($d_m$) | 256 | 256 | 256 | 256 |
| Dropout | 0.6 | 0.6 | 0.6 | 0.6 |
| Weight Decay | $5 \times 10^{-4}$ | $5 \times 10^{-4}$ | $5 \times 10^{-4}$ | $5 \times 10^{-4}$ |
| Epochs | 50 | 50 | 50 | 50 |
| Optimizer | AdamW | AdamW | AdamW | AdamW |
| *Cost-Sensitive Loss ($\mathcal{L}_{3WD}$) Parameters* | | | | |
| Polarity Misclassification Cost | 3.5 | 3.5 | 3.5 | 3.5 |
| Boundary Region Cost | 2.0 | 2.0 | 2.0 | 2.0 |
| Correct Prediction Weight | 0.2 | 0.2 | 0.2 | 0.2 |

- **Scale & Granularity:** CH-SIMSv2 [36] is significantly larger (4403 vs. 2281 samples) and features more fine-grained sentiment intensity annotations (5 classes: strongly negative, weakly negative, neutral, weakly positive, strongly positive) compared to CH-SIMS's [1] simpler 3-class scheme (negative, neutral, positive). This imposes a greater challenge in discerning subtle emotional differences.
- **Modality Alignment Focus:** A core objective of CH-SIMSv2 was to explicitly improve upon the *modality alignment quality* observed in CH-SIMS. The creators implemented stricter annotation protocols and quality control measures to ensure better synchronization and relevance between the visual/acoustic cues and the corresponding text/sentiment label [36]. Consequently, CH-SIMSv2 serves as a more robust benchmark for evaluating a model's ability to handle real-world modality misalignment.
- **Diversity:** CH-SIMSv2 sources clips from a wider variety of genres and platforms, aiming for greater naturalness and diversity compared to the initial CH-SIMS collection.

Consistent with established practices (Hazarika et al. [16]; Yu et al. [17]; Zhang et al. [14]), we adopt different metrics for each dataset: CH-SIMS and CH-SIMSv2 employ 3-class (Acc-3) and 5-class accuracy (Acc-5), while MOSI and MOSEI use 7-class accuracy (Acc-7). All datasets report 2-class accuracy (Acc-2), Mean Absolute Error (MAE), Pearson Correlation (Corr), and F1-score (F1). For MOSI and MOSEI, Acc-2 and F1 are computed under two classification schemes: negative vs. non-negative (has-0) and negative vs. positive (non-0). Lower MAE indicates better performance, whereas all other metrics improve with higher values.

### 4.2. Experimental settings

Our proposed 3WD-DRT framework was implemented and evaluated on four widely used public benchmarks: CH-SIMS, CH-SIMSv2, MOSI, and MOSEI. All experiments were conducted on a single NVIDIA RTX 3090 GPU to ensure a consistent computational environment. For feature extraction, we adopted BERT encoders tailored to the text modality, using `bert-base-chinese` for the Chinese datasets (CH-SIMS, CH-SIMSv2) and `bert-base-uncased` for the English datasets (MOSI, MOSEI). For the visual and acoustic modalities, we directly utilized the features provided with the original datasets.

The training process follows a decoupled two-stage strategy: a unimodal knowledge-enhancement pre-training stage, followed by an end-to-end multimodal fine-tuning stage. A key component of our framework is the cost-sensitive loss function, $\mathcal{L}_{3WD}$, which applies asymmetric penalties to different types of prediction errors. Specifically, we assign a high cost (3.5) for severe polarity misclassifications, a moderate cost (2.0) for boundary–region errors, and predictions within the correct polarity are assigned a low weight of 0.2, as detailed in Table 4.

We employed the AdamW optimizer for model training. The initial learning rate was set to $3 \times 10^{-5}$ or $4 \times 10^{-5}$ depending on the dataset. Batch sizes were set to 32 or 64, accordingly. A comprehensive list of all hyperparameter configurations is detailed in Table 4.

### 4.3. Baselines

To assess the effectiveness of 3WD-DRT, we compare it with a set of widely adopted baselines that fall into two primary categories. The first includes representation fusion models that focus on integrating multimodal features, such as TFN [38], LMF [19], MuLT [2],

**Table 5**
Performance comparison of our proposed 3WD-DRT model with baseline methods on the CH-SIMS and CH-SIMSv2 datasets. All metrics are reported as percentages (%) except for MAE and Corr. For MAE, lower values indicate better performance, while for all other metrics, higher values are better. The best results for each metric are highlighted in **bold**.

| Method | CH-SIMS | | | | | | CH-SIMSv2 | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | MAE | Corr | Acc-5 | Acc-3 | Acc-2 | F1 | MAE | Corr | Acc-5 | Acc-3 | Acc-2 | F1 |
| TFN | 0.432 | 0.591 | 39.30 | 65.12 | 78.38 | 78.62 | 0.303 | 0.707 | 52.55 | 72.21 | 80.14 | 80.14 |
| LMF | 0.441 | 0.576 | 40.53 | 64.68 | 77.77 | 77.88 | 0.367 | 0.557 | 47.79 | 64.90 | 74.18 | 73.88 |
| MulT | 0.453 | 0.564 | 37.94 | 64.77 | 78.56 | 79.66 | 0.291 | 0.738 | 54.81 | 73.19 | 80.68 | 80.73 |
| BBFN | 0.430 | 0.564 | 40.92 | 61.05 | 78.12 | 77.88 | 0.300 | 0.708 | 53.29 | 71.47 | 78.53 | 78.41 |
| Self-MM | 0.425 | 0.595 | 41.53 | 65.47 | 80.04 | 80.44 | 0.311 | 0.695 | 52.77 | 72.61 | 79.69 | 79.76 |
| CubeMLP | 0.419 | 0.593 | 41.79 | 65.86 | 77.68 | 77.59 | 0.334 | 0.648 | 52.90 | 71.95 | 78.53 | 78.53 |
| CENet | 0.471 | 0.534 | 33.92 | 62.58 | 77.90 | 77.53 | 0.310 | 0.699 | 53.04 | 73.10 | 79.56 | 79.63 |
| TETFN | 0.420 | 0.577 | 41.79 | 63.24 | 81.18 | 80.24 | 0.310 | 0.695 | 54.47 | 73.65 | 79.73 | 79.81 |
| ALMT | **0.408** | 0.594 | 43.11 | 65.86 | 78.77 | 78.71 | 0.308 | 0.700 | 52.90 | 71.86 | 79.59 | 79.51 |
| TMBL | 0.429 | 0.592 | 41.58 | 65.43 | 79.12 | 78.75 | 0.313 | 0.706 | 52.03 | 73.02 | 80.46 | 80.36 |
| KuDA | **0.408** | 0.613 | **43.54** | 66.52 | 80.74 | 80.71 | **0.271** | 0.759 | 61.22 | **76.21** | 82.11 | 82.04 |
| 3WD-DRT | **0.408** | **0.627** | 43.44 | **66.64** | **81.37** | **81.91** | 0.276 | **0.764** | **64.20** | 75.86 | **83.05** | **84.01** |

**Table 6**
Performance comparison with baseline methods on the MOSI and MOSEI datasets. For the Acc-2 and F1 metrics, we report results for both non-binary/binary sentiment classification. Lower MAE indicates better performance, while higher values are better for all other metrics. The best results are marked in **bold**.

| Method | MOSI | | | | | MOSEI | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | MAE | Corr | Acc-7 | Acc-2 | F1 | MAE | Corr | Acc-7 | Acc-2 | F1 |
| TFN | 0.947 | 0.673 | 34.46 | 77.99/79.08 | 77.95/79.11 | 0.572 | 0.714 | 51.60 | 78.50/81.89 | 78.96/81.74 |
| LMF | 0.950 | 0.651 | 33.82 | 77.90/79.18 | 77.80/79.15 | 0.576 | 0.717 | 51.59 | 80.54/83.48 | 80.94/83.36 |
| MulT | 0.879 | 0.702 | 36.91 | 79.71/80.98 | 79.63/80.95 | 0.559 | 0.733 | 52.84 | 81.15/84.63 | 81.56/84.52 |
| MISA | 0.776 | 0.778 | 41.37 | 81.84/83.54 | 81.82/83.58 | 0.557 | 0.751 | 52.05 | 80.67/84.67 | 81.12/84.66 |
| BBFN | 0.796 | 0.744 | 43.88 | 80.32/82.47 | 80.21/82.44 | 0.545 | 0.760 | 52.88 | 82.87/85.73 | 83.13/85.56 |
| MMIM | 0.744 | 0.780 | 44.75 | 82.51/84.30 | 82.38/84.23 | 0.550 | 0.761 | 51.88 | 83.75/85.42 | 83.93/85.26 |
| Self-MM | 0.708 | **0.796** | 46.67 | 83.44/85.46 | 83.36/85.43 | 0.531 | 0.764 | **53.87** | 83.76/85.15 | 83.82/84.90 |
| CubeMLP | 0.755 | 0.772 | 43.44 | 80.76/82.32 | 81.77/84.23 | 0.537 | 0.761 | 53.35 | 82.36/85.23 | 82.61/85.04 |
| CLGSI | 0.703 | 0.790 | 47.96 | 83.97/86.43 | 83.63/86.25 | 0.532 | 0.763 | 54.56 | 84.01/86.32 | **84.21**/86.18 |
| ALMT | 0.712 | 0.793 | 46.79 | 83.97/85.82 | 84.05/85.86 | 0.530 | 0.774 | 53.62 | 81.54/85.99 | 81.05/86.05 |
| KuDA | 0.705 | 0.795 | 47.08 | **84.40**/86.43 | 84.48/86.46 | 0.529 | 0.776 | 52.89 | 83.26/86.46 | 82.97/86.59 |
| 3WD-DRT | **0.701** | 0.794 | **48.42** | 83.41/**86.51** | **85.28/87.08** | **0.518** | **0.779** | 53.83 | **84.19/87.33** | 83.48/**88.26** |

MISA [16], SelfMM [17], CubeMLP [39], and TMBL [40]. The second category consists of guided fusion approaches, which aim to balance or adaptively guide information from dominant modalities. This group includes methods addressing text modality bias—such as MMIM [27], BBFN [41], CENet [42], TETFN [43], and ALMT [14]—as well as dynamic guidance strategies like KuDA [18] and approaches based on CLIP [44].

### 4.4. Main result

We empirically evaluate the proposed 3WD-DRT model on four widely-used multimodal sentiment analysis benchmarks: CH-SIMS, CH-SIMSv2, MOSI, and MOSEI. As shown in Tables 5 and 6, 3WD-DRT consistently achieves superior performance across diverse evaluation metrics, outperforming a broad range of baseline methods. Figs. 3 and 4 provide a visual comparison that highlights the model's comprehensive advantages.

On the Chinese datasets CH-SIMS and CH-SIMSv2, 3WD-DRT demonstrates clear improvements over prior approaches. Specifically, on CH-SIMS, it attains the highest Pearson Correlation (0.627), the best Acc-2 (81.37 %), and a top F1-Score (81.91 %), while matching the best MAE (0.408). Compared with strong competitors such as KuDA, 3WD-DRT shows advantages in F1-Score (81.91 % vs. 80.71 %) and Acc-3 (66.64 % vs. 66.52 %). On the more complex CH-SIMSv2 dataset, 3WD-DRT further improves performance, achieving an F1-Score of 84.01 %, the highest Corr (0.764), and the best results on Acc-5 (64.20 %) and Acc-2 (83.05 %), surpassing models like Self-MM. The radar plots in Fig. 3 confirm the model's balanced and robust performance.

On the English benchmarks MOSI and MOSEI, 3WD-DRT also achieves leading results. On MOSI, it records the lowest MAE (0.701), the highest Acc-7 (48.42 %), and a new best F1-Score (87.08 %), exceeding baselines like KuDA (86.46 %) and ALMT (85.86 %). Similarly, on MOSEI, 3WD-DRT ranks first in four of five major metrics: MAE (0.518), Corr (0.779), Acc-2 (87.33 %), and F1-Score (88.26 %), outperforming strong baselines such as Self-MM, particularly in MAE and F1. These findings are further illustrated in Fig. 4, where 3WD-DRT exhibits consistent, high-quality results.

Collectively, these results across multiple datasets substantiate the effectiveness and robustness of 3WD-DRT, confirming its state-of-the-art performance in multimodal sentiment analysis.

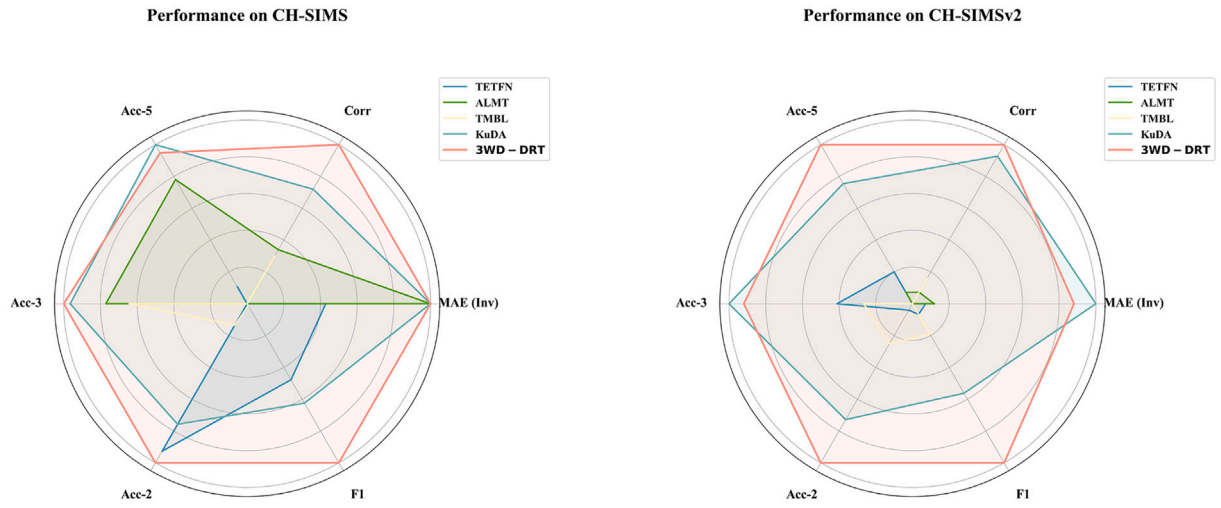


**Fig. 3.** Radar chart visualization of performance comparison on the CH-SIMS (left) and CH-SIMSv2 (right) datasets. The charts compare our 3WD-DRT model against top-performing baselines across six key metrics. To ensure a larger area consistently corresponds to better performance, the MAE metric has been inverted (labeled as MAE (Inv)). Our model, 3WD-DRT (represented by the shaded red area), consistently encloses the largest area, signifying its superior and more balanced overall performance compared to all baselines.

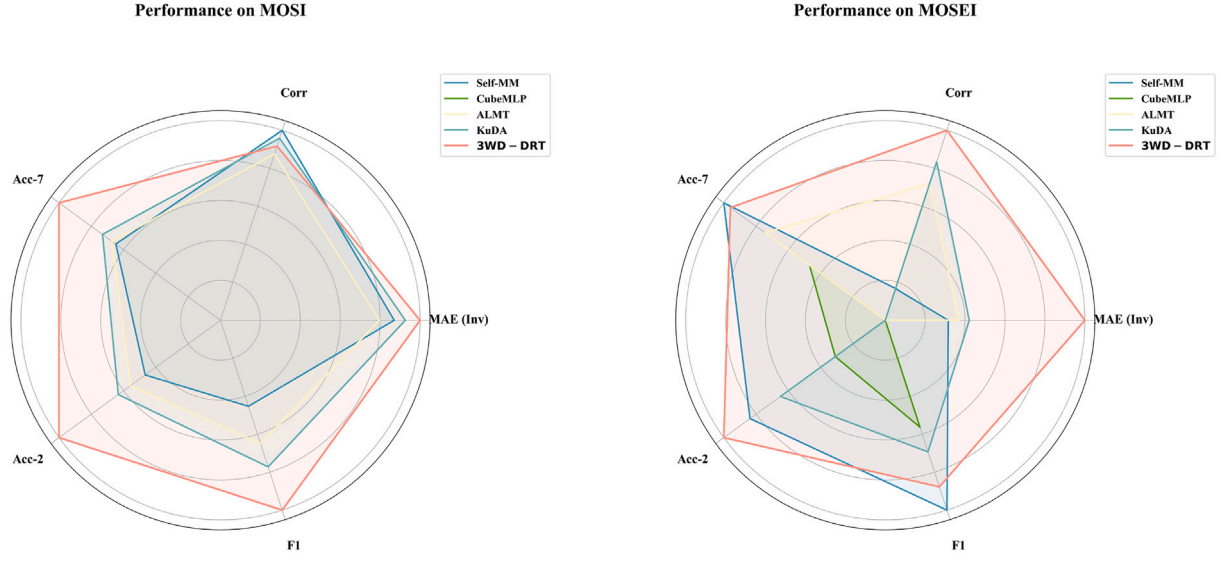


**Fig. 4.** Radar chart visualization of model performance on the MOSI (left) and MOSEI (right) datasets. The charts illustrate the comprehensive advantages of our 3WD-DRT model compared to competitive baselines across five key evaluation metrics. For visualization purposes, the MAE metric is inverted (labeled as MAE (Inv)) so that higher values reflect better performance, meaning a larger enclosed area consistently indicates superior overall performance. As shown, our 3WD-DRT model (represented by the shaded red area) consistently encloses the largest or one of the largest areas on both datasets. This visually confirms its state-of-the-art performance, particularly highlighting its leading results in the critical F1-Score and MAE metrics.

### *4.5. Statistical significance analysis*

To validate the robustness and reliability of the observed performance improvements of our proposed 3WD-DRT model over strong baselines, we conducted formal statistical significance tests following established practices [45–47]. Each model was independently trained and evaluated over 5 runs with different random seeds, generating paired samples of performance metrics on each dataset. Such repeated experiments allow capturing the natural variability of model performance and enable rigorous statistical comparison. We applied two complementary tests to assess the significance of improvements:

- **Wilcoxon signed-rank test:** A non-parametric test suitable for paired comparisons without assuming normal distribution of metric differences.
- **Corrected paired two-tailed *t*-test:** A parametric test assessing mean differences while accounting for potential dependencies between runs.

**Table 7**
Statistical significance test results comparing 3WD-DRT with KuDA [18] on key metrics for CH-SIMS and CH-SIMSv2 datasets. † and ‡ indicate $p < 0.05$ and $p < 0.01$ respectively.

| Metric | CH-SIMS | | | | | | CH-SIMSv2 | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | MAE | Corr | Acc-5 | Acc-3 | Acc-2 | F1 | MAE | Corr | Acc-5 | Acc-3 | Acc-2 | F1 |
| Wilcoxon | † | † | † | † | † | ‡ | † | † | † | † | † | ‡ |
| Paired *t*-test | † | † | † | † | † | ‡ | † | † | † | † | † | ‡ |

**Table 8**
Statistical significance test results comparing 3WD-DRT with KuDA [18] on key metrics for MOSI and MOSEI datasets. † and ‡ indicate $p < 0.05$ and $p < 0.01$ respectively.

| Metric | MOSI | | | | | MOSEI | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | MAE | Corr | Acc-7 | Acc-2 | F1 | MAE | Corr | Acc-7 | Acc-2 | F1 |
| Wilcoxon | † | † | † | † | ‡ | † | † | † | † | ‡ |
| Paired *t*-test | † | † | † | † | ‡ | † | † | † | † | ‡ |

Both tests were applied to compare 3WD-DRT against the strongest baseline KuDA across key metrics on all four datasets.

Tables 7 and 8 summarizes the p-values obtained from these tests. We mark statistically significant improvements with † for $p < 0.05$ and highly significant ones with ‡ for $p < 0.01$.

The results consistently indicate that the improvements of 3WD-DRT over KuDA [18] are statistically significant across the majority of metrics and datasets, thereby reinforcing the validity and generalizability of our approach.

### 4.6. Ablation study and analysis

To evaluate the contribution of each component in our proposed architecture, we conducted a comprehensive ablation study. The full model was systematically simplified by removing one module at a time, allowing for a quantitative assessment of each resulting variant's performance. This process enables the measurement of the specific importance of each component to the model's overall effectiveness.

The experimental setup uses our proposed **Full Model** as the performance baseline. We designed six distinct variants, each defined by the exclusion of a single module: (A) **w/o Pre-training**, which initializes the model with random weights; (B) **w/o** $K_{dom}$, which omits the knowledge alignment component; (C) **w/o** $\mathcal{L}_{KA}$, which removes the knowledge alignment loss; (D) **w/o 3WD Gating**, which ablates the core three-way discriminative gating mechanism; (E) **w/o** $\mathcal{L}_{TD-attn}$, which excludes the regularization loss term for stabilizing dynamic fusion gating thresholds; and (F) **w/o Cost-sensitive Loss**, which uses a standard loss function that does not account for class imbalance. These variants were benchmarked against the Full Model on the CH-SIMS, CH-SIMSv2, MOSI, and MOSEI datasets.

The results, presented in Tables 9 and 10, clearly demonstrate the effectiveness of our design. A key finding is that removing any single component leads to a consistent degradation in performance across all metrics and datasets. This provides empirical validation that the modules are not redundant but contribute collectively to the model's performance. This finding is visually supported by the bar chart in Fig. 5, where the performance of each ablation variant consistently falls below the baseline set by the Full Model.

A closer analysis of the results reveals a distinct hierarchy in component importance. The most significant performance drops are consistently observed in variants (A) w/o Pre-training and (D) w/o 3WD Gating. On the MOSEI dataset, for instance, removing pre-trained weights (A) increases the MAE by 10.2 % (from 0.518 to 0.571) and decreases the F1-score by approximately 5.6 %. Likewise, removing the 3WD Gating (D) causes a substantial decline, with MAE increasing to 0.555 and Corr dropping to 0.748. These results highlight the foundational nature of both the initialized knowledge from pre-training and the feature selection enabled

**Table 9**
Ablation study results on the CH-SIMS and CH-SIMSv2 datasets. The "Full Model" consistently outperforms all variants where a single component is removed, empirically validating the contribution of each module.

| Method | CH-SIMS | | | | | | CH-SIMSv2 | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | MAE | Corr | Acc-5 | Acc-3 | Acc-2 | F1 | MAE | Corr | Acc-5 | Acc-3 | Acc-2 | F1 |
| Full Model | 0.408 | 0.627 | 43.44 | 66.64 | 81.37 | 81.91 | 0.276 | 0.764 | 64.20 | 75.86 | 83.05 | 84.01 |
| w/o Pre-training (A) | 0.432 | 0.583 | 38.05 | 62.11 | 78.21 | 78.15 | 0.358 | 0.710 | 58.91 | 71.33 | 79.83 | 79.77 |
| w/o $K_{dom}$ (B) | 0.424 | 0.598 | 39.12 | 63.85 | 79.52 | 79.40 | 0.347 | 0.728 | 60.75 | 72.89 | 80.95 | 81.02 |
| w/o $\mathcal{L}_{KA}$ (C) | 0.419 | 0.609 | 40.03 | 64.92 | 80.11 | 80.25 | 0.341 | 0.741 | 62.13 | 73.98 | 81.76 | 82.13 |
| w/o 3WD Gating (D) | 0.426 | 0.595 | 38.88 | 63.02 | 79.13 | 79.01 | 0.351 | 0.721 | 59.84 | 72.10 | 80.44 | 80.59 |
| w/o $\mathcal{L}_{TD-attn}$ (E) | 0.410 | 0.624 | 41.28 | 66.41 | 81.21 | 81.53 | 0.329 | 0.761 | 63.92 | 75.61 | 82.81 | 83.56 |
| w/o Cost-sensitive Loss (F) | 0.418 | 0.612 | 39.57 | 64.33 | 79.85 | 79.72 | 0.340 | 0.745 | 61.59 | 73.16 | 81.33 | 81.48 |

**Table 10**
Ablation study results on the MOSI and MOSEI datasets. Similar to findings on the SIMS dataset, removing any single component leads to a degradation in performance, underscoring the necessity of each part of the proposed architecture.

| Method | MOSI | | | | | MOSEI | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | MAE | Corr | Acc-7 | Acc-2 | F1 | MAE | Corr | Acc-7 | Acc-2 | F1 |
| Full Model | 0.701 | 0.794 | 48.42 | 83.41/86.51 | 85.28/87.08 | 0.518 | 0.779 | 53.83 | 84.19/87.33 | 83.48/88.26 |
| w/o Pre-training (A) | 0.769 | 0.723 | 42.15 | 78.52/81.98 | 80.11/82.31 | 0.571 | 0.734 | 48.22 | 79.81/83.05 | 79.13/82.99 |
| w/o $K_{dom}$ (B) | 0.745 | 0.739 | 44.03 | 80.11/83.42 | 81.98/83.75 | 0.552 | 0.751 | 49.98 | 81.02/84.78 | 80.89/84.81 |
| w/o $\mathcal{L}_{KA}$ (C) | 0.728 | 0.748 | 45.89 | 81.67/84.99 | 83.15/84.88 | 0.540 | 0.763 | 51.06 | 82.76/85.91 | 82.11/86.13 |
| w/o 3WD Gating (D) | 0.748 | 0.735 | 43.61 | 79.88/83.01 | 81.12/83.41 | 0.555 | 0.748 | 49.53 | 80.79/84.52 | 80.05/84.40 |
| w/o $\mathcal{L}_{TD-attn}$ (E) | 0.704 | 0.759 | 48.11 | 83.12/86.23 | 84.88/86.65 | 0.520 | 0.777 | 52.81 | 83.92/87.15 | 83.21/87.88 |
| w/o Cost-sensitive Loss (F) | 0.733 | 0.741 | 44.98 | 80.65/84.02 | 82.23/84.15 | 0.545 | 0.759 | 50.78 | 81.91/85.26 | 81.35/85.39 |

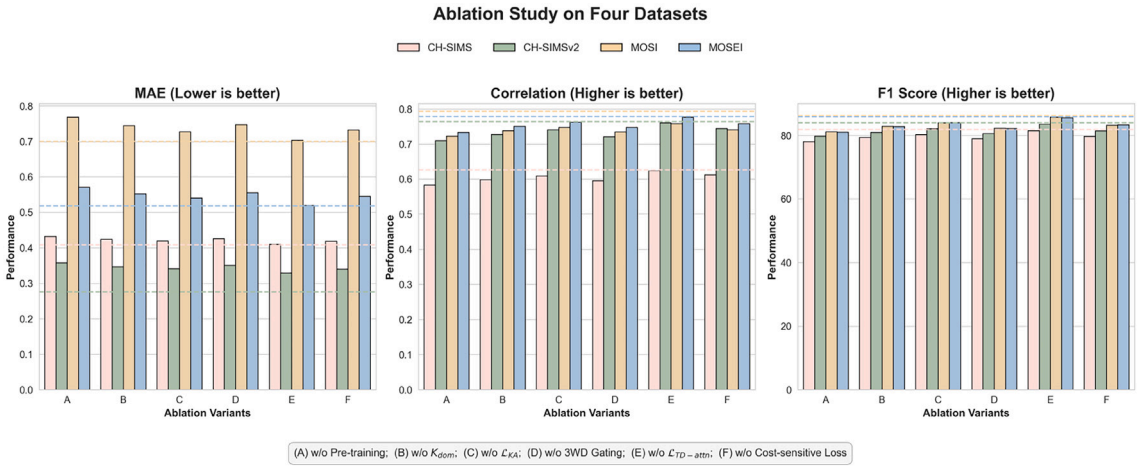


**Fig. 5.** Visual comparison of model performance in the ablation study. The bars represent the absolute scores of each ablation variant (A-F) across three key metrics. (A) w/o Pre-training, (B) w/o $K_{dom}$, (C) w/o $\mathcal{L}_{KA}$, (D) w/o 3WD Gating, (E) w/o $\mathcal{L}_{TD-attn}$, (F) w/o Cost-sensitive Loss. The corresponding dashed lines indicate the superior performance of the Full Model, visually demonstrating the performance gap caused by removing any single component.

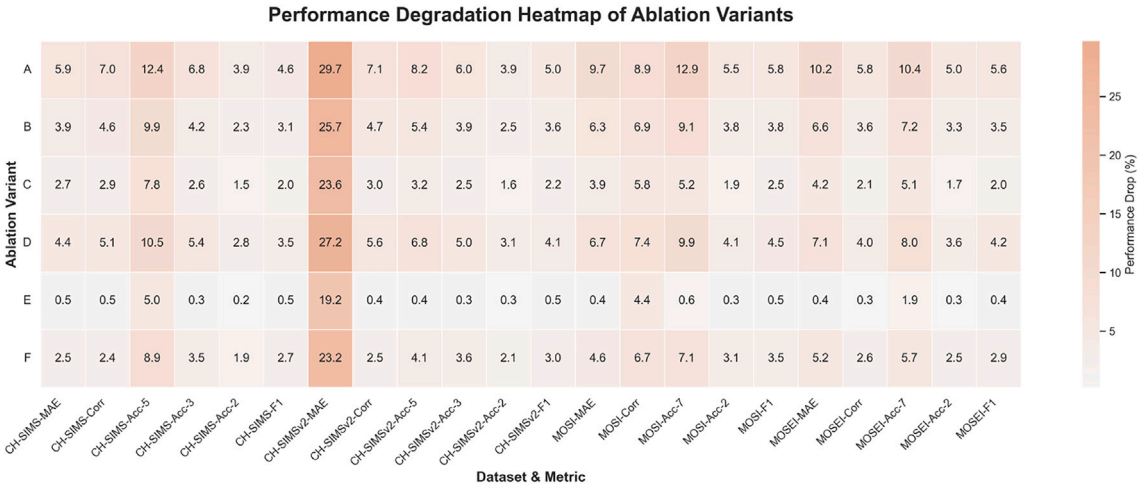


**Fig. 6.** Heatmap quantifying the percentage of performance degradation for each ablation variant relative to the Full Model. Warmer colors indicate a more substantial performance loss, thereby highlighting the most critical components. Consistent with the findings in Fig. 5, variants (A) w/o Pre-training and (D) w/o 3WD Gating consistently show the warmest colors, quantitatively identifying them as the most impactful modules in our architecture.

by the gating mechanism. The heatmap in Fig. 6 quantifies this relationship, where warmer colors correspond to larger performance drops, thereby identifying the most impactful components.

The contributions of the remaining modules, while less pronounced, are also statistically and practically significant. The removal of the knowledge alignment (B), the alignment loss (C), or the cost-sensitive loss (F) leads to tangible performance drops, which confirms

**Table 11**
Sensitivity results for cost-sensitive loss hyperparameters on the MOSI and MOSEI datasets. For each parameter, the table reports F1-Score (%) and MAE at different tested values. Default configurations are underlined, and optimal settings are in bold. The results indicate that PMC and BRC peak at moderate levels, while CPW should remain small for best performance.

| Hyperparameter | Value | MOSI | | MOSEI | |
|---|---|---|---|---|---|
| | | F1 | MAE | F1 | MAE |
| Polarity Misclassification Cost(BRC = 2.0, CPW = 0.2) | 2.5 | 86.85 % | 0.706 | 88.02 % | 0.523 |
| | 3 | 86.99 % | 0.703 | 88.17 % | 0.52 |
| | **3.5** | **87.08 %** | **0.701** | **88.26 %** | **0.518** |
| | 4 | 87.05 % | 0.702 | 88.21 % | 0.519 |
| | 4.5 | 86.91 % | 0.705 | 88.10 % | 0.522 |
| Boundary Region Cost(PMC = 3.5, CPW = 0.2) | 1 | 86.81 % | 0.707 | 88.05 % | 0.524 |
| | 1.5 | 86.95 % | 0.703 | 88.18 % | 0.52 |
| | **2** | **87.08 %** | **0.701** | **88.26 %** | **0.518** |
| | 2.5 | 86.92 % | 0.704 | 88.15 % | 0.521 |
| | 3 | 86.77 % | 0.708 | 88.01 % | 0.525 |
| Correct Prediction Weight(PMC = 3.5, BRC = 2.0) | 0.05 | 87.01 % | 0.702 | 88.20 % | 0.519 |
| | 0.1 | 87.05 % | 0.701 | 88.24 % | 0.518 |
| | **0.2** | **87.08 %** | **0.701** | **88.26 %** | **0.518** |
| | 0.5 | 86.75 % | 0.709 | 87.95 % | 0.526 |
| | 1 | 86.42 % | 0.715 | 87.61 % | 0.531 |

their roles in refining the model's predictive accuracy and robustness. Interestingly, variant (E) w/o $\mathcal{L}_{TD-attn}$ shows the smallest performance degradation; on the CH-SIMS dataset, its MAE increases by only 0.002. Nevertheless, the Full Model still outperforms this variant on every metric. This indicates that the target-driven attention mechanism, while subtle in its effect, provides a necessary refinement to achieve optimal performance.

Collectively, these results support our architectural choices. The consistent performance degradation observed when any single component is removed confirms that all modules are integral to the final outcome. The interplay between foundational elements, such as pre-training and gating, and components for refinement, like attention and specialized losses, is evidently essential for the model's success.

### 4.7. Sensitivity analysis of cost-sensitive loss hyperparameters

A key element of our framework is the cost-sensitive loss $\mathcal{L}_{3WD}$, which uses a predefined cost matrix to impose asymmetric penalties on different types of prediction errors. To examine the robustness of the selected weights and assess their effect on model performance, we performed a sensitivity analysis on two English multimodal sentiment datasets, MOSI and MOSEI. In each experiment, one hyperparameter was varied while the other two were fixed at their optimal settings (Polarity Misclassification Cost = 3.5, Boundary Region Cost = 2.0, Correct Prediction Weight = 0.2). The corresponding results are summarized in Table 11 and illustrated in Fig. 7, where the line plots show the performance trends for each hyperparameter in terms of both F1-score and Mean Absolute Error (MAE).

The Polarity Misclassification Cost (PMC) emerges as the most influential hyperparameter, heavily penalizing severe semantic errors. As PMC increases from 2.5 to 3.5, the F1-score for both datasets rises, peaking at 3.5. This indicates that a sufficiently high penalty effectively drives the model to avoid polarity errors. Beyond this point (4.0 and 4.5), performance declines slightly, suggesting that an excessive penalty may destabilize training or encourage overly conservative predictions, thereby limiting generalization.

For the Boundary Region Cost (BRC), which penalizes predictions in uncertain regions, optimal performance is observed at a moderate value of 2.0. Smaller costs (e.g., 1.0 or 1.5) fail to adequately discourage ambiguous predictions, while larger costs (e.g., 2.5 or 3.0) tend to over-penalize subtle or neutral cases, both leading to marginal reductions in the F1-score.

The Correct Prediction Weight (CPW) assigns a small penalty to correct predictions, ensuring that all samples contribute to the loss. The best performance occurs at a CPW of 0.2. When this weight is increased to 0.5 or 1.0, the model's F1-score drops, as the relative gap between penalties for correct and incorrect predictions narrows, weakening the cost-sensitive emphasis of the loss.

Overall, the results in Table 11 confirm that PMC and BRC each have a clear mid-range optimum, while CPW is most effective when kept small. Fig. 7 complements these results by showing the non-linear performance curves, where F1-score (left axis) and MAE (right axis) are plotted together. The dual-axis layout allows direct visual comparison of classification accuracy and regression error trends across varying penalty weights.

### 4.8. Efficiency and parameter analysis

To assess the practical viability of the proposed 3WD-DRT model for real-world deployment, we analyzed its computational efficiency, parameter count, and inference cost.

First, we compared our model's efficiency with baseline model, KuDA, as both are built upon a BERT-base architecture. The results, shown in Table 12, indicate that our proposed three-way decision mechanism adds only a marginal overhead. With an increase of approximately 2 M parameters (1.6 %), the average inference time increases by less than 1 ms. This demonstrates the cost-effectiveness of our approach, as it achieves significant performance gains with only a marginal increase in computational overhead.

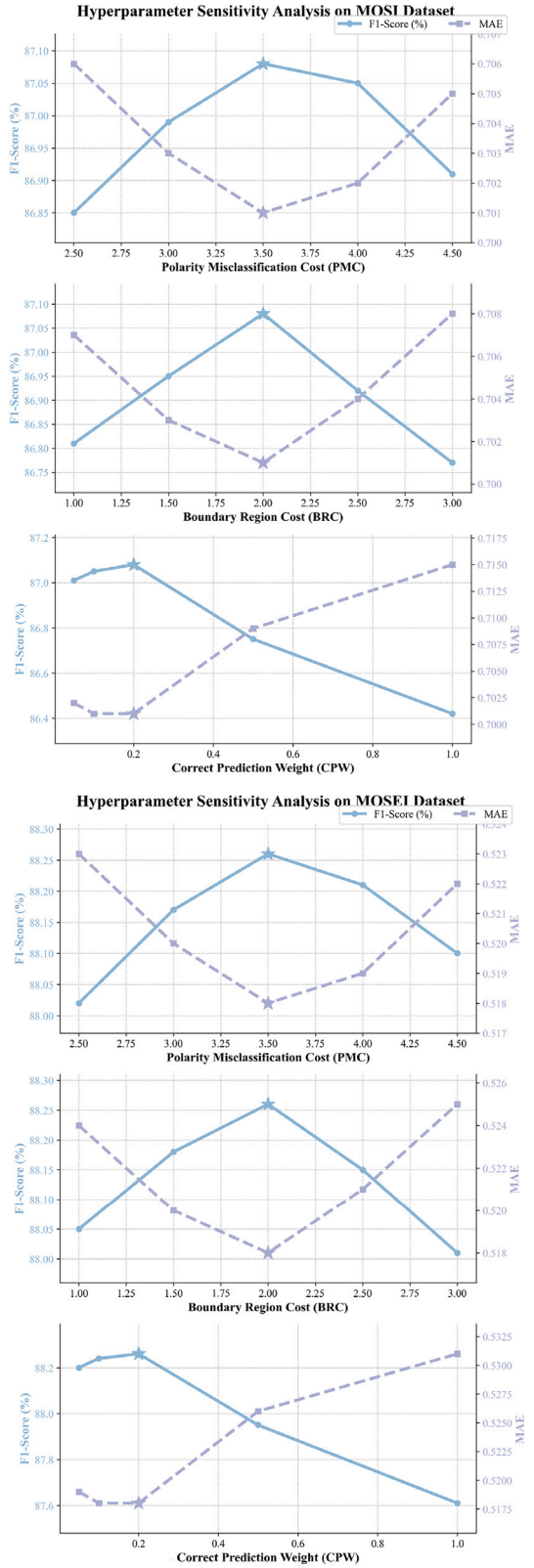


**Fig. 7.** Sensitivity analysis of cost-sensitive loss hyperparameters on the MOSI and MOSEI datasets. Each row of subplots corresponds to one hyperparameter: Polarity Misclassification Cost (PMC), Boundary Region Cost (BRC), and Correct Prediction Weight (CPW). Blue solid lines with circular markers represent F1-Score (%, left axis), and purple dashed lines with square markers represent MAE (right axis). Star markers indicate optimal hyperparameter values. The plots clearly show PMC's benefit up to a certain threshold, BRC's peak at a moderate level, and CPW's performance drop when set too high.

**Table 12**
Efficiency comparison with the KuDA baseline.

| Model | Total parameters | Avg. inference time |
|---|---|---|
| KuDA | 124.77 M | 37.42 ms/batch |
| **3WD-DRT** | **126.83 M** | **38.17 ms/batch** |

**Table 13**
Performance and parameter trade-off on MOSI by replacing the text backbone.

| Method | MOSI | | | | | parameter |
|---|---|---|---|---|---|---|
| | MAE | Corr | Acc-7 | Acc-2 | F1 | |
| 3WD-DRT(bert-base-uncased) | 0.701 | 0.794 | 48.42 | 83.41/86.51 | 85.28/87.08 | 126.83 M |
| 3WD-DRT(TinyBERT ) | 0.732 | 0.775 | 46.510 | 82.15/85.20 | 83.90/85.95 | 31.36 M |

**Table 14**
Performance evaluation of the 3WD-DRT model under simulated modality failure on the CH-SIMS dataset. T, V, and A represent Text, Vision, and Audio, respectively. For MAE, lower is better; for all other metrics, higher is better.

| Modality Combination | CH-SIMS | | | | | |
|---|---|---|---|---|---|---|
| | MAE | Corr | Acc-5 | Acc-3 | Acc-2 | F1 |
| T + V + A (Full) | 0.408 | 0.627 | 43.44 | 66.64 | 81.37 | 81.91 |
| T + V | 0.413 | 0.619 | 40.95 | 66.01 | 81.02 | 81.35 |
| T + A | 0.415 | 0.616 | 40.78 | 65.83 | 80.91 | 81.17 |
| V + A | 0.449 | 0.551 | 36.88 | 60.95 | 76.84 | 76.99 |
| T-only | 0.421 | 0.607 | 40.15 | 65.12 | 80.24 | 80.33 |
| V-only | 0.498 | 0.482 | 31.02 | 55.43 | 70.15 | 70.28 |
| A-only | 0.512 | 0.466 | 29.86 | 54.19 | 69.53 | 69.41 |

Furthermore, we investigated the impact of the text backbone on the model's overall efficiency and performance. The majority of the model's parameters ( 110 M) reside in the bert-base-uncased encoder. We conducted an experiment on the MOSI dataset where we replaced this backbone with the lightweight TinyBERT (approx. 14.5 M parameters). As shown in Table 13, this experiment revealed a clear trade-off between performance and efficiency.

### *4.9. Robustness to modality failure*

To evaluate the resilience of our proposed 3WD-DRT model against partial data loss—a common real-world scenario—we conducted a series of experiments simulating modality failure. For these tests, we simulated the failure of a modality (Text, Vision, or Audio) by zeroing out its entire input feature sequence. This method was intentionally chosen to model a complete data–channel loss (e.g., a video without an audio track) and to rigorously test how our dynamic routing mechanism adapts to such a drastic absence of information. While other techniques like feature-level dropout test for general robustness to noise, our approach provides a clear assessment of the model's performance under conditions of complete modality unavailability, which directly evaluates the effectiveness of our proposed three-way decision gate. The experiments were conducted on the CH-SIMS dataset.

As detailed in Table 14 and illustrated in Fig. 8, the 3WD-DRT model demonstrates substantial robustness. The absence of any single modality leads to a predictable drop in performance compared to the full trimodal model, which set a benchmark F1 score of 81.91 % and a correlation of 0.627. The magnitude of this performance degradation is directly proportional to the relative importance of the omitted modality.

Further analysis highlights the pivotal role of text. The most significant performance decline occurs when the textual modality is removed (the V + A combination), where the F1 score falls to 76.99 %. In contrast, removing the visual (T + A) or auditory (T + V) modalities results in a milder reduction, yielding F1 scores of 81.17 % and 81.35 %, respectively. This trend is reinforced by the unimodal results; the text-only model achieves a strong F1 score of 80.33 %, notably surpassing the performance of the visual-only (70.28 %) and audio-only (69.41 %) configurations.

The model does not suffer a catastrophic breakdown with incomplete data but instead exhibits graceful degradation. Combinations that retain the text modality maintain a high level of performance, which underscores the model's ability to leverage the most informative available inputs. This behavior is not an artifact but rather direct evidence of our dynamic routing fusion mechanism's effectiveness. The mechanism intelligently identifies modality loss and recalibrates the fusion process to prioritize reliable inputs, thereby ensuring stable performance under imperfect data conditions.

### *4.10. Case study*

Fig. 9 depicts the reasoning process of our 3WD-DRT model in a challenging case from the CH-SIMS dataset. The core challenge is resolving the emotional conflict and ambiguity among modalities, a task that requires intelligent judgment rather than a simple fusion of information.

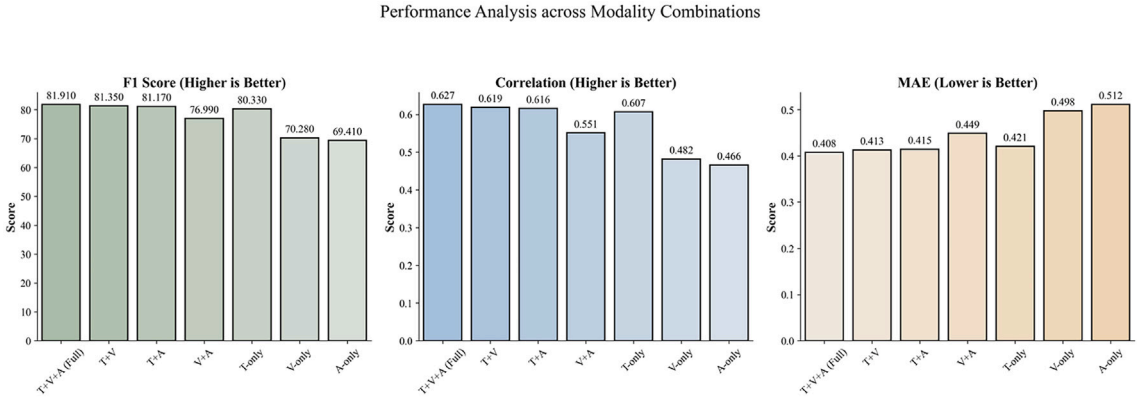


**Fig. 8.** Visualization of key performance metrics (F1 Score, Correlation, and MAE) across different modality combinations. The chart illustrates the model's graceful degradation as modalities are removed, highlighting the central role of the Text (T) modality in maintaining high performance.

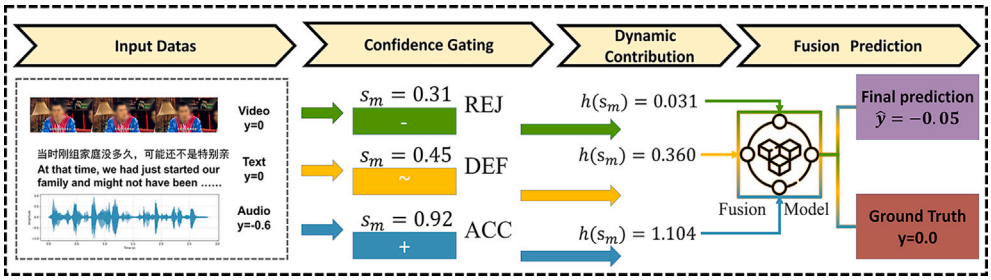


**Fig. 9.** This case illustrating the reasoning process of our 3WD-DRT model on a challenging case. The figure details a four-stage process from initial signal evaluation to the final fused prediction. Visual attributes such as opacity and flowline thickness represent the model's internal states: modality confidence ($s_m$) and dynamic contribution ($h(s_m)$), respectively.

**Stage 1: Input Signals** The input signals present a subtle conflict: both the Text ($y_T = 0.0$) and Visual ($y_V = 0.0$) modalities are emotionally neutral, while the Audio ($y_A = -0.6$) conveys a slightly negative sentiment. The sample's overall multimodal ground truth ($y_M = 0.0$), however, is neutral, which demands that the model accurately reconcile this inconsistency instead of being misled by the single negative signal.

Stage 2: Confidence Gating, As depicted, the model first assesses the discriminative value of each modality to assign a confidence score ($s_m$). Rather than weighting the most emotionally intense modality highest, the model assigns the greatest confidence to the Audio modality ($s_m^A = 0.92$), identifying it as the sole source of non-neutral information. Consequently, the audio signal is routed to the "Acceptance" (ACC) region for amplification. In contrast, the neutral Text and Visual modalities receive lower confidence scores ($s_m^T = 0.45$ and $s_m^V = 0.31$) and are mapped to the "Deferment" (DEF) and "Rejection" (REJ) regions for corresponding modulation and suppression.

Stage 3: Dynamic Contribution, This gating decision directly controls each modality's dynamic contribution to the final fusion, a process visually represented by the thickness of the flowlines in the diagram. The high confidence in the Audio signal results in its contribution being substantially amplified with a dynamic scaling factor of $h_{s_m}^A = 1.104$. The Text signal's contribution is moderately adjusted ($h_{s_m}^T = 0.360$), while the low-confidence Visual signal is almost entirely suppressed ($h_{s_m}^V = 0.031$).

Stage 4: Fused Prediction, The model's final prediction is not a simple weighted average but a non-linear fusion of these dynamically scaled contributions. A significantly amplified negative signal from the audio is balanced against the two suppressed neutral signals, yielding a highly precise prediction of $\hat{y} = -0.05$. This result is remarkably close to the ground– truth ($y = 0.0$), while still capturing the subtle negative sentiment from the audio.

This case demonstrates the model's ability to go beyond simple fusion. The underlying mechanism can identify the most discriminative modality from conflicting information and dynamically adjust its influence, enabling robust and precise final predictions.

## 5. Conclusion

A fundamental challenge in artificial intelligence is teaching machines to accurately interpret the complex and often contradictory emotional signals from language, facial expressions, and vocal tones. Prevailing models in MSA often struggle for two primary reasons. First, they rely on rigid, static methods for fusing these data streams and fail to adapt when one modality, such as a sarcastic tone of voice, overrides the literal meaning of the text. Second, their learning is guided by traditional loss functions that treat all mistakes as

equal, overlooking the fact that misinterpreting a positive emotion as negative is a far more significant error than a minor misjudgment of emotional intensity.

This paper introduces the 3WD-DRT, a novel framework that addresses these limitations through a "quality-aware, decision-driven" paradigm. Instead of simply combining all inputs, the model first dynamically assesses the confidence and reliability of each modality—vision, audio, and text—for any given sample. Based on this assessment, a three-way decision gate adaptively amplifies the most informative signals while modulating or suppressing those deemed noisy or misleading. Additionally, we designed a novel cost-sensitive loss function that imposes greater penalties for severe semantic errors, such as polarity misclassifications, aligning the model's learning process more closely with human perception of error.

Extensive experiments on four widely-used benchmark datasets (CH-SIMS, CH-SIMSv2, MOSI, and MOSEI) validate our framework. The 3WD-DRT model consistently outperformed a broad range of state-of-the-art methods across all datasets and evaluation metrics. For instance, on the large-scale MOSEI dataset, our model set a new benchmark with a MAE of 0.518 and an F1-Score of 88.26 %. Ablation studies confirmed that each component of our architecture is integral to its success, with the dynamic gating mechanism and pre-training stages being particularly critical. The model also showed remarkable robustness and exhibited graceful performance degradation when entire modalities of data were missing, validating the effectiveness of its dynamic fusion mechanism.

Beyond experimental validation, the proposed framework has significant real-world potential. In human–computer interaction, it can power virtual assistants and social robots to respond with appropriate empathy, improving user engagement and trust. In healthcare, especially mental health monitoring, it can assist clinicians in detecting subtle emotional cues from patients' multimodal signals, aiding early diagnosis of depression or anxiety. In public opinion monitoring, it can analyze large-scale multimedia content to capture shifts in societal sentiment, enabling timely policy or communication responses. These applications illustrate that 3WD-DRT not only advances the state of the art in MSA but also lays a foundation for deploying emotionally intelligent AI in diverse and impactful domains.

In addressing these core challenges in MSA, this work provides a theoretically grounded and empirically validated solution that moves toward more nuanced and robust AI systems.

While the current study demonstrates the effectiveness and versatility of 3WD-DRT, several directions remain open for future exploration. First, applying the framework to real-time multimodal streams, where modality reliability may change dynamically with latency and noise, would test its adaptability under practical deployment constraints. Second, integrating self-supervised or weakly supervised learning strategies could reduce the reliance on large-scale annotated datasets, making the approach more scalable and applicable to low-resource domains. Third, the decision-driven design philosophy may be extended to related affective computing tasks beyond sentiment analysis, such as empathy recognition, multimodal deception detection, or cognitive workload estimation. Exploring these directions could further broaden the impact and applicability of the proposed paradigm.

## CRediT authorship contribution statement

**Haoyu Jiang:** Writing – original draft, Software, Formal analysis, Data curation, Conceptualization. **Xiaoliang Chen:** Writing – review & editing, Supervision, Project administration, Methodology, Investigation, Funding acquisition. **Duoqian Miao:** Resources, Funding acquisition. **Hongyun Zhang:** Validation. **Xiaolin Qin:** Resources, Project administration, Funding acquisition. **Shangyi Du:** Visualization. **Peng Lu:** Writing – review & editing.

## Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Acknowledgment

This work is supported by the National Natural Science Foundation (Grant nos. 625762876, 2402395), the National Key R&D Program of China (Grant nos. 2023YFB3308601, 2022YFB3104700), Chengdu "Open bidding for selecting the best candidates" Science and Technology Project (Grant no. 2023-JB00-00020-GX), the Science and Technology Program of Sichuan Province (Grant no. 2023YFS0424), the Science and Technology Service Network Initiative (Grant no. KFJ-STS-QYZD-2021–21-001), and the Talents Program by Sichuan Provincial Party Committee Organization Department, and Chengdu - Chinese Academy of Sciences Science and Technology Cooperation Fund Project (Major Scientific and Technological Innovation Projects).

## Data availability

Data are available for download at the following web links: https://github.com/Joeisjoejoe/3WD.

## References

[1] W. Yu, H. Xu, F. Meng, Y. Zhu, Y. Ma, J. Wu, J. Zou, K. Yang, Ch-SIMS: a Chinese multimodal sentiment analysis dataset with fine-grained annotation of modality, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online, 2020, pp. 3718–3727, https://doi.org/10.18653/v1/2020.acl-main.343

[2] Y.-H.-H. Tsai, S. Bai, P.P. Liang, J.Z. Kolter, L.-P. Morency, R. Salakhutdinov, Multimodal transformer for unaligned multimodal language sequences, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Florence, Italy, 2019, pp. 6558–6569, https://doi.org/10.18653/v1/P19-1656

[3] P. Wang, Q. Zhou, Y. Wu, T. Chen, J. Hu, Dlf: disentangled-language-focused multimodal sentiment analysis, in: AAAI Conference on Artificial Intelligence, 2024, https://api.semanticscholar.org/CorpusID:274788853
[4] Y. Yao, Three-way decision: an interpretation of rules in rough set theory, in: Rough Sets and Knowledge Technology, Springer Berlin Heidelberg, Berlin, Heidelberg, 2009, pp. 642–649, https://doi.org/10.1007/978-3-642-02962-2_81
[5] Y. Yao, Three-way decision, three-world conception, and explainable AI, in: Rough Sets, Springer Nature Switzerland, Cham, 2022, pp. 39–53, https://doi.org/10.1007/978-3-031-21244-4_4
[6] S. Luo, Three-way decision in a multi-source information system and its applications, IEEE Access 7 (2019) 108343–108359, https://doi.org/10.1109/ACCESS.2019.2933259
[7] J. Ye, J. Zhan, B. Sun, A three-way decision method based on fuzzy rough set models under incomplete environments, Inf. Sci. 577 (2021) 22–48, https://doi.org/10.1016/j.ins.2021.06.088
[8] B. Zhou, Y. Yao, J. Luo, A three-way decision approach to eMail spam filtering, in: Advances in Artificial Intelligence, Springer Berlin Heidelberg, Berlin, Heidelberg, 2010, pp. 28–39, https://doi.org/10.1007/978-3-642-13059-5_6
[9] J. Chen, Y. Chen, Y. He, et al, A classified feature representation three-way decision model for sentiment analysis, Appl. Intell. 52 (7) (2022) 7995–8007, https://doi.org/10.1007/s10489-021-02809-1
[10] Z. Zhang, R. Wang, Applying three-way decisions to sentiment classification with sentiment uncertainty, in: Rough Sets and Knowledge Technology, Springer International Publishing, Cham, 2014, pp. 720–731, https://doi.org/10.1007/978-3-319-11740-9_66
[11] T. Wang, H. Li, L. Zhang, X. Zhou, B. Huang, A three-way decision model based on cumulative prospect theory, Inf. Sci. 519 (2020) 74–92, https://doi.org/10.1016/j.ins.2020.01.030
[12] J. Wang, X. Zhang, G. Yu, Y. Chen, S. Rao, Deebert-s3wd: Three-way multigranularity decision for interactive information sentiment analysis research, Math. Probl. Eng. 2022 (1) (2022) 1090777, https://doi.org/10.1155/2022/1090777
[13] J. Su, W. Liu, D. Feng, C. Shi, Y. Liu, Social media sentiment analysis of sequential three-way decision model based on knowledge fusion, in: 2023 IEEE 3rd International Conference on Electronic Technology, Communication and Information (ICETCI), 2023, pp. 20–24, https://doi.org/10.1109/ICETCI57876.2023.10176764
[14] H. Zhang, Y. Wang, G. Yin, K. Liu, Y. Liu, T. Yu, Learning language-guided adaptive hyper-modality representation for multimodal sentiment analysis, in: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Singapore, 2023, pp. 756–767, https://doi.org/10.18653/v1/2023.emnlp-main.49
[15] Q. Pan, Z. Meng, Hybrid uncertainty calibration for multimodal sentiment analysis, Electronics 13 (3) (2024) 662, https://doi.org/10.3390/electronics13030662
[16] D. Hazarika, R. Zimmermann, S. Poria, Misa: modality-invariant and -specific representations for multimodal sentiment analysis, in: Proceedings of the 28th ACM International Conference on Multimedia, Association for Computing Machinery, New York, NY, USA, 2020, pp. 1122–1131, https://doi.org/10.1145/3394171.3413678
[17] W. Yu, H. Xu, Z. Yuan, J. Wu, Learning modality-specific representations with self-supervised multi-task learning for multimodal sentiment analysis, in: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 35, AAAI Press, 2021, pp. 10790–10797, https://doi.org/10.1609/aaai.v35i12.17289
[18] X. Feng, Y. Lin, L. He, Y. Li, L. Chang, Y. Zhou, Knowledge-guided dynamic modality attention fusion framework for multimodal sentiment analysis, in: Findings of the Association for Computational Linguistics: EMNLP 2024, 2024, pp. 14755–14766, https://doi.org/10.18653/v1/2024.findings-emnlp.865
[19] Z. Liu, Y. Shen, V.B. Lakshminarasimhan, P.P. Liang, A. Bagher Zadeh, L.-P. Morency, Efficient low-rank multimodal fusion with modality-specific factors, in: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Melbourne, Australia, 2018, pp. 2247–2256, https://doi.org/10.18653/v1/P18-1209
[20] A. Bhat, R. Mahar, R. Punia, R. Srivastava, Exploring multimodal sentiment analysis through cartesian product approach using BERT embeddings and ResNet-50 encodings and comparing performance with pre-existing models, in: 2022 3rd International Conference for Emerging Technology (INCET), 2022, pp. 1–6, https://doi.org/10.1109/INCET54531.2022.9825245
[21] S. Sabour, N. Frosst, G.E. Hinton, Dynamic routing between capsules, ArXiv abs/1710.09829, 2017, https://api.semanticscholar.org/CorpusID:3603485
[22] S. Sun, G. Xu, S. Lu, Mfm: Multimodal sentiment analysis based on modal focusing model, in: 2024 IEEE International Conference on Systems, Man, and Cybernetics (SMC), 2024, pp. 1524–1529, https://doi.org/10.1109/SMC54092.2024.10831023
[23] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, L.U. Kaiser, I. Polosukhin, Attention is all you need, in: Advances in Neural Information Processing Systems, vol. 30, Curran Associates, Inc., 2017, https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf
[24] A. Pandey, D.K. Vishwakarma, Progress, achievements, and challenges in multimodal sentiment analysis using deep learning: a survey, Appl. Soft Comput. 152 (C) (Feb. 2024) 111206, https://doi.org/10.1016/j.asoc.2023.111206
[25] B.D. Jieyu an, W.M.N.W. Zainon, Leveraging vision-language pre-trained model and contrastive learning for enhanced multimodal sentiment analysis, Intell. Autom. Soft Comput. 37 (2) (2023) 1673–1689, https://doi.org/10.32604/iasc.2023.039763
[26] M.B. Habib, M.F.B. Hafiz, N.A. Khan, S. Hossain, Multimodal sentiment analysis using deep learning fusion techniques and transformers, Int. J. Adv. Comput. Sci. Appl. 15 (6) (2024) https://doi.org/10.14569/IJACSA.2024.0150686
[27] W. Han, H. Chen, S. Poria, Improving multimodal fusion with hierarchical mutual information maximization for multimodal sentiment analysis, in: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 2021, pp. 9180–9192, https://doi.org/10.18653/v1/2021.emnlp-main.723
[28] A. Amini, W. Schwarting, A. Soleimany, D. Rus, Deep evidential regression, in: Advances in Neural Information Processing Systems, vol. 33, Curran Associates, Inc., 2020, pp. 14927–14937, https://proceedings.neurips.cc/paper_files/paper/2020/file/aab085461de182608ee9f607f3f7d18f-Paper.pdf
[29] Y. Gal, Z. Ghahramani, Dropout as a Bayesian approximation: representing model uncertainty in deep learning, in: Proceedings of the 33rd International Conference on Machine Learning, Vol. 48 of Proceedings of Machine Learning Research, PMLR, New York, New York, USA, 2016, pp. 1050–1059, https://proceedings.mlr.press/v48/gal16.html
[30] N. Thai-Nghe, Z. Gantner, L. Schmidt-Thieme, Cost-sensitive learning methods for imbalanced data, in: The 2010 International Joint Conference on Neural Networks (IJCNN), 2010, pp. 1–8, https://doi.org/10.1109/IJCNN.2010.5596486
[31] Z. Qin, C. Zhang, T. Wang, S. Zhang, Cost sensitive classification in data mining, in: Advanced Data Mining and Applications, Springer Berlin Heidelberg, 2010, pp. 1–11, https://doi.org/10.1007/978-3-642-17316-5_1
[32] H. Zhang, L. Jiang, C. Li, Cs-resnet: Cost-sensitive residual convolutional neural network for PCB cosmetic defect detection, Expert Syst. Appl. 185 (2021) 115673, https://doi.org/10.1016/j.eswa.2021.115673
[33] B.N. Njike, X. Siebert, Nonparametric active learning for cost-sensitive classification, CoRR abs/2310.00511, 2023, https://doi.org/10.48550/arXiv.2310.00511
[34] W. Zheng, H. Zhao, Cost-sensitive hierarchical classification for imbalance classes, Appl. Intell. 50 (8) (2020) 2328–2338, https://doi.org/10.1007/S10489-019-01624-Z
[35] A. van den Oord, Y. Li, O. Vinyals, Representation learning with contrastive predictive coding, ArXiv abs/1807.03748, 2018, https://api.semanticscholar.org/CorpusID:49670925
[36] Y. Liu, Z. Yuan, H. Mao, Z. Liang, W. Yang, Y. Qiu, T. Cheng, X. Li, H. Xu, K. Gao, Make acoustic and visual cues matter: CH-SIMS v2.0 dataset and AV-Mixup consistent module, in: Proceedings of the 24th ACM International Conference on Multimodal Interaction (ICMI '22), ACM, Bengaluru, India, 2022, pp. 1–12, https://doi.org/10.1145/3536221.3556630
[37] A. Zadeh, R. Zellers, E. Pincus, L.-P. Morency, Multimodal sentiment intensity analysis in videos: facial gestures and verbal messages, IEEE Intell. Syst. 31 (6) (2016) 82–88, https://doi.org/10.1109/MIS.2016.94
[38] A. Bagher Zadeh, P.P. Liang, S. Poria, E. Cambria, L.-P. Morency, Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph, in: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Melbourne, Australia, 2018, pp. 2236–2246, https://doi.org/10.18653/v1/P18-1208

[39] H. Sun, H. Wang, J. Liu, Y. Chen, L. Lin, Cubemlp: a MLP-based model for multimodal sentiment analysis and depression estimation, CoRR abs/2207.14087, 2022, https://doi.org/10.48550/ARXIV.2207.14087
[40] J. Huang, J. Zhou, Z. Tang, J. Lin, C.Y.-C. Chen, Tmbl: Transformer-based multimodal binding learning model for multimodal sentiment analysis, Knowl.-Based Syst. 285 (2024) 111346, https://doi.org/10.1016/j.knosys.2023.111346
[41] W. Han, H. Chen, A. Gelbukh, A. Zadeh, L.-P. Morency, S. Poria, Bi-bimodal modality fusion for correlation-controlled multimodal sentiment analysis, in: Proceedings of the 2021 International Conference on Multimodal Interaction, Association for Computing Machinery, New York, NY, USA, 2021, pp. 6–15, https://doi.org/10.1145/3462244.3479919
[42] D. Wang, S. Liu, Q. Wang, Y. Tian, L. He, X. Gao, Cross-modal enhancement network for multimodal sentiment analysis, IEEE Trans. Multimed. 25 (2023) 4909–4921, https://doi.org/10.1109/TMM.2022.3183830
[43] D. Wang, X. Guo, Y. Tian, J. Liu, L. He, X. Luo, Tetfn: a text enhanced transformer fusion network for multimodal sentiment analysis, Pattern Recognit. 136 (2023) 109259, https://doi.org/10.1016/j.patcog.2022.109259
[44] Y. Yang, X. Dong, Y. Qiang, CLGSI: A multimodal sentiment analysis framework based on contrastive learning guided by sentiment intensity, in: Findings of the Association for Computational Linguistics: NAACL 2024, Association for Computational Linguistics, Mexico City, Mexico, 2024, pp. 2099–2110, https://doi.org/10.18653/v1/2024.findings-naacl.135
[45] J. Demšar, Statistical comparisons of classifiers over multiple data sets, J. Mach. Learn. Res. 7 (2006) 1–30, https://dl.acm.org/doi/pdf/10.5555/1248547.1248548
[46] L. Jiang, L. Zhang, L. Yu, D. Wang, Class-specific attribute weighted naive Bayes, Pattern Recognit. 88 (2019) 321–330, https://doi.org/10.1016/j.patcog.2018.11.032
[47] L. Jiang, L. Zhang, C. Li, J. Wu, A correlation-based feature weighting filter for Naive Bayes, IEEE Trans. Knowl. Data Eng. 31 (2) (2019) 201–213, https://doi.org/10.1109/TKDE.2018.2836440