# Conflict-Aware Hyper-Modality for Multimodal Sentiment Analysis

**Anonymous EMNLP submission**

## Abstract

Multimodal Sentiment Analysis (MSA) often suffers from semantic conflicts among modalities, particularly in complex scenarios like sarcasm where textual semantics contradict visual or acoustic cues. Conventional methods typically assume that all modalities contribute constructively to a global consensus, leaving them highly vulnerable to semantic pollution during unregulated fusion. To address this critical challenge, we propose the Conflict-Aware Hyper-Modality Learning (CHML) framework, designed to explicitly perceive and model intermodal conflicts. CHML introduces learnable hyper-modality representations guided by hierarchical linguistic semantics and employs a Modality Conflict Gate (MCG) to dynamically suppress unreliable cross-modal interactions based on feature similarity. Furthermore, we decouple modality-specific features from the hyper-modality and mine deep complementary information via symmetric cross-attention. A language-guided gated fusion strategy is then adopted to preserve critical sentiment information while adapting to contextual variations. Extensive experiments on CH-SIMS, CMU-MOSI, and CMU-MOSEI demonstrate that CHML achieves state-of-the-art performance, and comprehensive ablation studies further verify the efficacy and robustness of our framework against semantic ambiguity.

**Keywords:** Multimodal Sentiment Analysis, Conflict-Aware Learning, Cross-Modal Fusion, Hyper-Modality Representation

## 1 Introduction

MSA achieves accurate emotion recognition by fusing heterogeneous signals, driving continuous advancements in human-computer interaction (**?**Zhao et al., 2025). By mining complementary cross-modal cues, MSA significantly outperforms unimodal approaches (Baltrušaitis et al., 2018), with recent Transformer-based architectures further propelling the field via fine-grained attention mechanisms (Vaswani et al., 2017). However, a critical bottleneck remains: auxiliary modalities often contain emotion-irrelevant noise, and more problematically, conflicting semantic information. Environmental variations and inherent semantic ambiguity severely limit model performance in complex scenarios like sarcasm (Dhanith et al., 2024) (Figure 1).

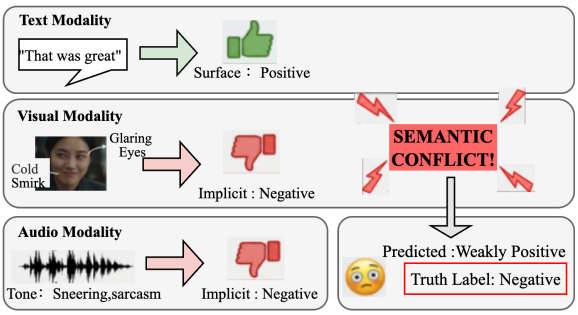



Figure 1: Examples of modality semantic conflicts in MSA.

To tackle the pressing challenge of semantic incongruity, recent studies have explored diverse paradigms, ranging from incongruity-aware graph networks to causal interventions (Wang et al., 2023a; Lin et al., 2023). It is widely acknowledged that modality contributions in MSA are inherently unbalanced, with the linguistic modality typically dominating (Hazarika et al., 2020). Motivated by this, researchers have leveraged text semantics to guide non-linguistic modalities, treating language as a "semantic anchor" to filter noise. For instance, ALMT (Zhang et al., 2023) utilizes multi-scale linguistic features to guide visual and acoustic representations, while CENet (Wang et al., 2022a) dynamically injects visual and acoustic information into pre-trained language models.

Despite these advancements, conventional shared-representation paradigms blindly assume that all modalities contribute constructively toward a global consensus (Wang et al., 2024). However,

in conflicting scenarios such as sarcasm, unregulated fusion inevitably triggers semantic pollution, causing the predominant modality (e.g., deceptive text) to completely overshadow subtle but crucial paralinguistic cues (Liu et al., 2023). Lacking dedicated mechanisms to explicitly detect and regulate these inter-modal conflicts, existing models are highly susceptible to biased or erroneous predictions (Chen et al., 2024a).

To bridge this gap, we propose the Conflict-Aware Hyper-Modality Learning (CHML) framework, designed to enhance MSA by explicitly mitigating the impact of conflicting information. Specifically, we introduce a Conflict-Adaptive Hyper-modality Learning (CAHL) module that leverages hierarchical linguistic semantics to generate hyper-modality representations. Within CAHL, a Modality Conflict Gate (MCG) is embedded to dynamically suppress unreliable cross-modal interactions by tracking the cosine similarity between linguistic and auxiliary modalities. Furthermore, we design a Feature Disentanglement and Cross-Modal Information Mining (CAMI) module to decouple pristine modality-specific features from the global consensus, mutually mining complementary cues via symmetric cross-attention. Finally, an adaptive language-guided gated fusion strategy integrates these refined features to achieve robust sentiment prediction. Our main contributions are threefold:

- We propose CHML, a novel framework that explicitly models modality conflicts by treating inter-modal inconsistency as an exploitable signal rather than mere noise.
- We design the Modality Conflict Gate (MCG) to quantify inter-modal discrepancies, generating adaptive gating weights that dynamically suppress conflicting modalities and prevent semantic pollution.
- We introduce a "Reconstruction-Interaction" strategy that decodes pristine modality-specific features directly from the hyper-modality, facilitating bidirectional interaction via symmetric cross-attention to ensure a high signal-to-noise ratio.

## 2 Related Work

### 2.1 Strategies for Multimodal Fusion

Effective heterogeneous information fusion remains the crux of MSA. Early strategies, ranging from heuristic concatenation to tensor-based fusion (e.g., TFN (Zadeh et al., 2017)), struggled with dynamic dependencies and the curse of dimensionality. Subsequently, attention-based architectures (e.g., Transformers (Vaswani et al., 2017), MulT (Tsai et al., 2019)) and graph networks (e.g., GCNet (Wang et al., 2023b)) significantly advanced fine-grained cross-modal interactions. However, these methods predominantly rely on passive weight adjustments, inherently assuming all interactions are constructive. Without explicit mechanisms to perceive and filter inter-modal conflicts, they remain highly vulnerable to deceptive signals in incongruous contexts.

### 2.2 Approaches for Handling Modality Inconsistency

Modality inconsistency is a critical bottleneck in MSA. Existing solutions primarily explore representation disentanglement, dynamic gating, and hierarchical guidance. Disentanglement approaches (e.g., MISA (Hazarika et al., 2020), DMD (Li et al., 2023a), SDRS (Zhao et al., 2025b)) isolate modality-specific features via reconstruction losses, but struggle with sensitive hyper-parameter tuning and information leakage (Chen et al., 2024b). To filter unaligned noise, uncertainty-aware routing algorithms (Gao et al., 2024a; Wu et al., 2025) estimate predictive confidence to mitigate conflicts. However, they typically quantify information volume rather than semantic discrepancies. Alternatively, hierarchical strategies (e.g., HTRN (Li et al., 2024a)) leverage linguistic anchors to purify non-textual features, though they prioritize alignment over conflict resolution, risking a text-centric bias. Furthermore, despite the cross-modal reasoning capabilities of recent Multimodal LLMs (Sun et al., 2025), their uniform input processing renders fine-grained conflict handling an uninterpretable black box (Li et al., 2024b). Ultimately, current paradigms lack dedicated mechanisms to explicitly detect and reconcile semantic conflicts under contradictory emotional cues (Zhang et al., 2024; Zhou et al., 2024).

## 3 Methodology

Figure 2 illustrates the overall pipeline of the proposed Conflict-Aware Hyper-Modality Learning (CHML) framework.

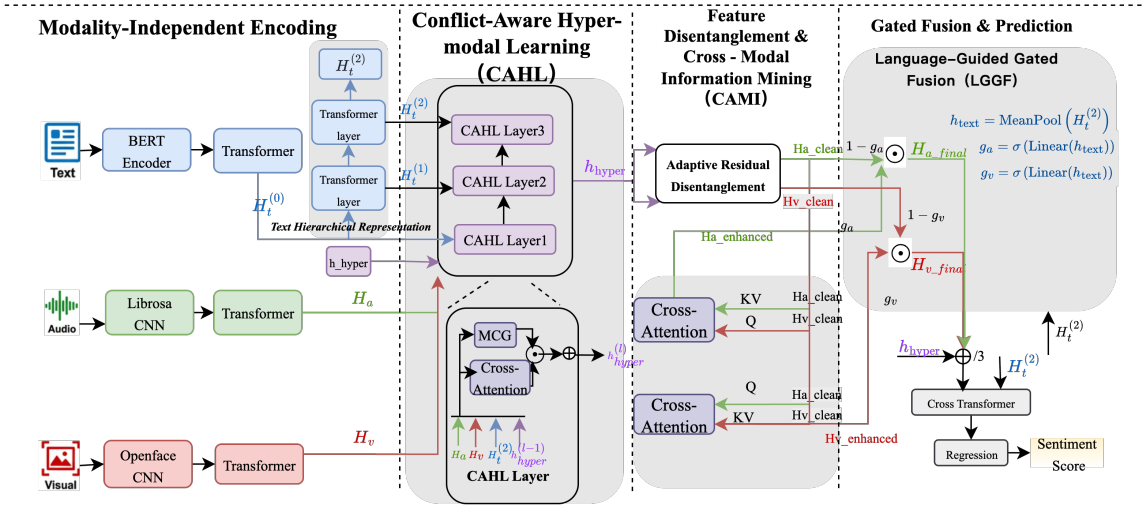



Figure 2: Overview of the CHML framework.

### 3.1 Modality-Independent Encoding

To map inputs into a unified semantic space, linguistic features are extracted via a pre-trained BERT. Visual and acoustic sequences are processed by a hybrid 1D-CNN (kernels 5 and 7) and a Transformer encoder to yield $H_m \in \mathbb{R}^{T_m \times d_h}$, where $m \in \{a, v\}$ and the hidden dimension is set to $d_h = 128$. Recognizing language's dominance, the base text feature $H_t^{(0)}$ is fed into a 2-layer Transformer to generate a hierarchical representation set $\{H_t^{(0)}, H_t^{(1)}, H_t^{(2)}\}$, capturing multi-level semantics. Concurrently, a trainable hyper-modality token $h_{\text{hyper}}$ is initialized to aggregate task-level global consensus.

### 3.2 Conflict-Aware Hyper-Modality Learning

The CAHL module explicitly tackles information redundancy and modality conflicts by iteratively learning a robust hyper-modality representation guided by hierarchical text features. At each layer $l$, it functions as follows (Figure 3).

**Modality Conflict Gate.** To prevent deceptive paralinguistic cues from polluting the shared consensus, MCG dynamically quantifies semantic conflicts by computing the token-level cosine similarity between textual and auxiliary features, preserving fine-grained discrepancies without losing temporal resolution:

$$\text{sim}_{tm}^{(l)} = \text{CosSim}\left(H_t^{(l-1)} W_t, H_m W_m\right).$$

Here, $m \in \{a, v\}$, where the similarity is calculated across the feature dimension at each time step, preserving temporal resolution instead of global pooling. The adaptive gate weight $g_t^{(l)}$ is then generated via a linear projection and a Sigmoid activation $\sigma$:

$$g_t^{(l)} = \sigma\left(W_{\text{gate}}\left[s_{ta}^{(l)} \| s_{tv}^{(l)}\right] + b_{\text{gate}}\right) \in (0, 1).$$

When a semantic conflict occurs (e.g., positive text with a sarcastic tone, resulting in low similarity at specific time steps), $g_t^{(l)}$ explicitly drops toward 0, suppressing misleading updates to the hyper-modality. Crucially, the raw acoustic and visual features $(H_a, H_v)$ remain strictly unmodified, ensuring unpolluted signals are forwarded to the downstream CAMI module.

**Language-Guided Cross-Attention Update.** To extract synergistic features, we project the modalities into a shared subspace and perform cross-modal attention, utilizing $H_t^{(l-1)}$ as the Query to attend to $H_a$ and $H_v$:

$$Q_t = H_t^{(l-1)} W_q,$$
$$K_m = H_m W_k, \quad V_m = H_m W_v.$$

$$Z_m^{(l)} = \text{Softmax}\left(\frac{Q_t K_m^\top}{\sqrt{d_h}}\right) V_m \in \mathbb{R}^{T_t \times d_h}.$$

To align this sequence-level shift with the token-level hyper-modality, we apply pooling and a Multi-Layer Perceptron, strictly modulated by the MCG gate $g_t^{(l)}$:

$$\text{shift}_t^{(l)} = \text{MLP}\left(\text{Pool}\left(Z_a^{(l)} + Z_v^{(l)}\right)\right),$$

$$h_{\text{hyper}}^{(l)} = h_{\text{hyper}}^{(l-1)} + g_t^{(l)} \cdot \text{shift}_t^{(l)}.$$

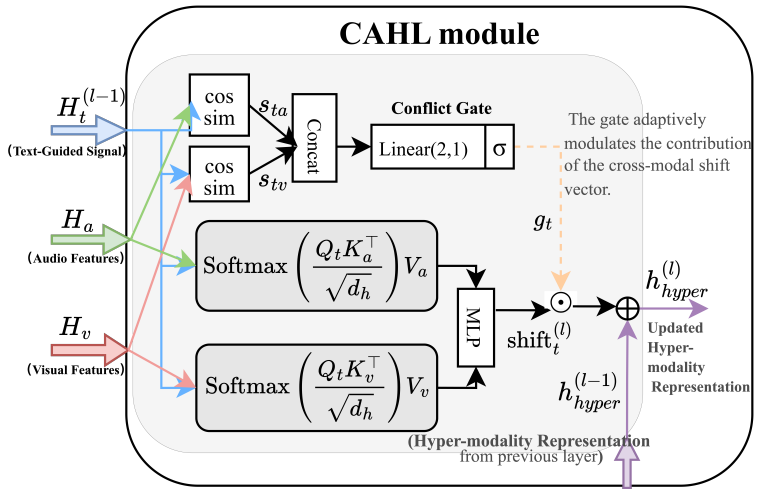



Figure 3: Overview of the CAHL module.

The final layer's output forms the robust global consensus $h_{\text{hyper}}^{\text{final}}$. This design enables CAHL to selectively absorb synergistic sentiment signals while rigorously shielding the consensus from ironic or ambiguous noise.

### 3.3 Feature Disentanglement and Interaction

Despite CAHL's efficacy in filtering semantic noise, it predominantly captures aligned semantics. Discriminative paralinguistic cues vital for resolving emotional ambiguities remain entangled in the original feature space. Therefore, CAMI is proposed to explicitly disentangle these modality-specific variations and bidirectionally mine complementary features.

**Adaptive Residual Disentanglement.** To isolate these variations, we project the hyper-modality back to modality-specific spaces to estimate the expected consensus $H_m^* = h_{\text{hyper}} W_m$. To prevent rigid hard subtraction from corrupting the intrinsic semantic structures of individual modalities, we introduce a learnable, sequence-wise adaptive scaling factor $\alpha \in \mathbb{R}^{T_m \times 1}$. This mechanism affords fine-grained temporal control over the disentanglement process, dynamically modulating the intensity of consensus filtering in response to the severity of cross-modal discrepancies (e.g., amplifying disentanglement during highly incongruous sarcastic contexts). Ultimately, we reliably reconstruct the idiosyncratic, modality-specific features by deducting the scaled global consensus from the raw representations, stabilized via Layer Normalization:

$$H_m^{\text{clean}} = \text{LayerNorm}\left(H_m - \alpha \odot H_m^*\right).$$

This residual disentanglement mathematically amplifies modality-specific polarities during semantic conflicts.

**Symmetric Cross-Attention Modality Interaction.** Subsequently, to mutually verify and integrate robust non-verbal cues, we bidirectionally mine complementary information via a Symmetric Cross-Attention mechanism:

$$H_m^{\text{enhanced}} = \text{CMHA}\big(Q = H_m^{\text{clean}}, \\ K = H_{\bar{m}}^{\text{clean}},\ V = H_{\bar{m}}^{\text{clean}}\big),$$

where $\bar{m} \in \{a, v\} \setminus \{m\}$ denotes the complementary auxiliary modality. This symmetric architecture ensures modality equality and a high signal-to-noise ratio.

### 3.4 Language-Guided Gated Fusion and Prediction

Ultimately, this stage leverages the overarching linguistic context to dynamically balance and fuse the pristine modality-specific cues with the enhanced representations, deriving the final global prediction.

**Language-Guided Gated Fusion (LGGF).** To adaptively integrate features, we derive a global text anchor:

$$h_{\text{text}} = \text{MeanPool}\left(H_t^{(2)}\right).$$

This anchor guides two independent gating networks to dynamically balance the pristine ($H_m^{\text{clean}}$) and enhanced ($H_m^{\text{enhanced}}$) cues:

$$H_m^{\text{final}} = g_m \odot H_m^{\text{enhanced}} \\ + (1 - g_m) \odot H_m^{\text{clean}},$$

where $g_m = \sigma\left(\text{Linear}(h_{\text{text}})\right)$ for $m \in \{a, v\}$.

**Cross-Transformer Fusion and Regression.** To achieve global consistency, the hyper-modality undergoes a recursive update by integrating it with fused paralinguistic features re-projected into the hyper-latent space. An averaging operation is employed to normalize the representation and suppress feature scale explosion:

$$h_{\text{hyper}}^{\text{final}} = \frac{1}{3}\Big(h_{\text{hyper}} + P_a\left(H_a^{\text{final}}\right) + P_v\left(H_v^{\text{final}}\right)\Big),$$

where $P_a$ and $P_v$ are linear projection layers. This cyclic refinement stabilizes training dynamics and enhances representation discriminability.

Finally, a Cross-Transformer captures deep inter-modal dependencies by attending the linguistic anchor $H_t^{(2)}$ (Query) to the global consensus $h_{\text{hyper}}^{\text{final}}$ (Key and Value):

$$h_{\text{out}} = \text{CrossTrans}\big(Q = H_t^{(2)}, \\ K = h_{\text{hyper}}^{\text{final}},\ V = h_{\text{hyper}}^{\text{final}}\big).$$

The output [CLS] embedding $h_{\text{out}}$ is fed into a linear regression layer to predict the sentiment intensity $\hat{y}$:

$$\hat{y} = W_r h_{\text{out}} + b_r.$$

The model is trained via Smooth L1 Loss (Huber Loss) to mitigate the impact of severe label noise inherent in sarcastic samples and ensure robust optimization:

$$\mathcal{L} = \frac{1}{N}\sum_{i=1}^{N} \text{Smooth}_{L1}(y_i - \hat{y}_i),$$

where $y_i$ is the ground-truth sentiment label.

# 4 Experiments

## 4.1 Experiment Settings

**Datasets.** We evaluate CHML on three benchmarks: (1) **CH-SIMS** (Wang et al., 2020): A Chinese dataset (2,281 clips) with dense semantic conflicts and sarcasm, ideal for evaluating conflict-aware models. (2) **CMU-MOSI** (Zadeh et al., 2018a) and (3) **CMU-MOSEI** (Zadeh et al., 2018b): Classical English datasets (2,199 and 22,856 clips, respectively) annotated with continuous sentiment intensities from $-3$ to $+3$.

**Baselines.** We extensively compare CHML against multiple state-of-the-art baselines: 1) Early/Tensor Fusion: TFN (Zadeh et al., 2017), LMF (Liu et al., 2018); 2) Alignment & Translation: MulT (Tsai et al., 2019), MFM (Singh and Kose, 2020); 3) Disentanglement: MISA (Hazarika et al., 2020), PMR (Lv et al., 2021); 4) Pre-trained & Transformer-based: MAG-BERT (Li et al., 2022), Self-MM (Guo et al., 2021), ALMT (Zhang et al., 2023); 5) Enhancement-based: MMIM (Cai et al., 2022), FDMER (Yang et al., 2022), CHFN (Wang et al., 2022b).

## 4.2 Comparison with State-of-the-Art Methods

As summarized in Table 1 and Table 2, CHML establishes new state-of-the-art performance across all benchmarks. To ensure a fair comparison, the results of baselines marked with † are directly adopted from the public MMSA benchmark framework. On the English MOSI and MOSEI datasets, CHML outperforms the highly competitive ALMT baseline (e.g., $+0.5\%$ Acc-2 and $+0.4\%$ F1-score on MOSI) and yields a significantly lower MAE than prior translation or disentanglement methods (e.g., MulT, MISA). This demonstrates that dynamically gating conflicts provides far more precise continuous predictions than unregulated fusion. Crucially, on the conflict-heavy CH-SIMS dataset, CHML achieves $80.96\%$ in Acc-2 (outperforming ALMT by nearly $0.9\%$) and maintains consistent superiority across fine-grained multi-class metrics. These substantial gains strongly validate our core motivation: explicitly quantifying and suppressing inter-modal inconsistencies via the MCG is pivotal for uncovering true emotional states in deceptive contexts.

## 4.3 Ablation Study and Analysis

**Effects of Different Components.** To validate the individual contributions within our framework, we design five variants on CH-SIMS: 1) w/o MCG: removes the Modality Conflict Gate; 2) w/o ARD: replaces Adaptive Residual Disentanglement with hard subtraction; 3) w/o CAMI: removes Symmetric Cross-Attention; 4) w/o LGGF: replaces Language-Guided Gated Fusion with vanilla concatenation; and 5) Baseline: directly concatenates initial features.

As shown in Table 3, the full CHML framework consistently outperforms all variants. We observe that: 1) *Necessity of Conflict Perception*: Removing MCG causes the most severe degradation ($-1.37\%$ in Acc-2). This dramatic decline proves that unregulated cross-modal attention introduces severe contradictory noise in sarcastic scenarios, highlighting MCG as a crucial shield. 2)

Table 1: Main results on MOSI and MOSEI. Best results are highlighted in bold. * represents the result is from Hazarika et al. (2020); † represents results adopted from the MMSA benchmark[1].

| Method | MOSI | | | | | | MOSEI | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Acc-2 | Acc-5 | Acc-7 | F1 | MAE | Corr | Acc-2 | Acc-5 | Acc-7 | F1 | MAE | Corr |
| TFN* | -/80.8 | - | 34.9 | -/80.7 | 0.901 | 0.698 | -/82.5 | - | 50.2 | -/82.1 | 0.593 | 0.700 |
| LMF* | -/82.5 | - | 33.2 | -/82.4 | 0.917 | 0.695 | -/82.5 | - | 48.0 | -/82.1 | 0.623 | 0.677 |
| MFM* | -/81.7 | - | 35.4 | -/81.6 | 0.877 | 0.706 | -/84.4 | - | 51.3 | -/84.3 | 0.568 | 0.717 |
| MulT | -/83.0 | - | 40.0 | -/82.8 | 0.871 | 0.698 | -/82.5 | - | 51.8 | -/82.3 | 0.580 | 0.703 |
| MISA | 81.8/83.4 | - | 42.3 | 81.7/83.6 | 0.783 | 0.761 | 83.6/85.5 | - | 52.2 | 83.8/85.3 | 0.555 | 0.756 |
| PMR | -/83.6 | - | 40.6 | -/83.4 | - | - | -/83.3 | - | 52.5 | -/82.6 | - | - |
| MAG-BERT | 82.3/84.4 | - | 43.62 | 82.5/84.6 | 0.727 | 0.781 | 82.5/84.8 | - | 52.6 | 82.7/84.7 | 0.543 | 0.755 |
| Self-MM | 82.5/84.7 | - | 45.79 | 83.6/84.9 | 0.712 | 0.795 | 82.6/84.9 | - | 53.4 | 82.9/84.9 | 0.529 | 0.767 |
| MMIM | 84.1/86.0 | - | 46.65 | 84.0/85.9 | 0.700 | 0.800 | 82.2/85.9 | - | 54.2 | 82.6/85.9 | 0.526 | 0.772 |
| FDMER | -/84.6 | - | 44.1 | -/84.7 | 0.724 | 0.788 | -/86.1 | - | 54.1 | -/85.8 | 0.536 | 0.773 |
| CHFN | 84.3/86.4 | - | **48.6** | 84.2/86.2 | 0.689 | 0.809 | 83.7/86.2 | - | 54.3 | 83.9/86.1 | **0.525** | 0.778 |
| ALMT† | 84.1/86.1 | **54.37** | 47.9 | 84.0/85.9 | 0.704 | 0.800 | 83.8/86.4 | 55.57 | 53.8 | 83.0/85.3 | 0.527 | 0.775 |
| **CHML** | **84.6/86.7** | 54.21 | 48.3 | **84.4/86.5** | **0.683** | **0.811** | **84.2/86.7** | **55.64** | **54.6** | **83.9/86.4** | **0.525** | **0.781** |

Table 2: Results on the Chinese dataset CH-SIMS. Note: the best result is highlighted in bold; † represents results adopted from the MMSA benchmark[1].

| Method | Acc-5 | Acc-3 | Acc-2 | F1 | MAE | Corr |
|---|---|---|---|---|---|---|
| TFN† | 39.3 | 65.12 | 78.38 | 78.62 | 0.432 | 0.591 |
| LMF† | 40.53 | 64.68 | 77.77 | 77.88 | 0.441 | 0.576 |
| MFM† | - | - | 75.06 | 75.58 | 0.477 | 0.525 |
| MulT† | 37.94 | 64.77 | 78.56 | 79.66 | 0.453 | 0.564 |
| MISA† | - | - | 76.54 | 76.59 | 0.447 | 0.563 |
| MAG-BERT† | - | - | 74.44 | 71.75 | 0.492 | 0.399 |
| Self-MM† | 41.53 | 65.47 | 80.04 | 80.44 | 0.425 | 0.595 |
| ALMT† | 43.10 | **66.73** | 80.08 | 79.89 | 0.413 | 0.595 |
| **CHML** | **43.98** | 66.52 | **80.96** | **80.59** | **0.410** | **0.620** |

Table 3: Ablation results on CH-SIMS.

| Architecture | Acc-2 | F1 | MAE ↓ | Corr ↑ |
|---|---|---|---|---|
| CHML | **80.96** | **80.59** | **0.410** | **0.620** |
| w/o ARD | 80.08 | 79.48 | 0.434 | 0.575 |
| w/o LGGF | 79.82 | 79.35 | 0.438 | 0.558 |
| w/o CAMI | 79.65 | 79.12 | 0.442 | 0.551 |
| w/o MCG | 79.15 | 78.5 | 0.456 | 0.540 |
| Baseline | 78.3 | 78.1 | 0.485 | 0.512 |

*Superiority of Adaptive Disentanglement*: While hard subtraction (w/o ARD) explicitly strips consensus, it aggressively damages modality-specific topological structures. Our adaptive residual factor effectively dynamically ensures optimal pristine feature reconstruction. 3) *Importance of Interaction & Dynamic Fusion*: The 0.87% Acc-2 drop in w/o CAMI validates that symmetric bidirectional interaction is vital for mining complementary nonverbal cues. Furthermore, the decline in w/o LGGF confirms that static fusion fails in context-varying ambiguities, whereas our language-guided gating successfully leverages textual anchors to adaptively weigh modal contributions.

Table 4: Effects of different fusion methods.

| Fusion method | CH-SIMS | | MOSI | |
|---|---|---|---|---|
| | Acc-5 | MAE | Acc-7 | MAE |
| Concatenation | 43.76 | 0.410 | 48.69 | 0.703 |
| GRU | 44.86 | 0.414 | 47.81 | 0.710 |
| Tensor Fusion | 44.20 | **0.403** | 47.23 | 0.710 |
| Low-Rank | 45.08 | 0.408 | 46.65 | 0.715 |
| Addition | 42.45 | 0.411 | 46.36 | 0.706 |
| Ours | **45.73** | 0.404 | **49.42** | **0.683** |

**Effects of Different Fusion Techniques.** To analyze the effects of different fusion techniques, we conduct experiments whose results are shown in Table 4. Obviously, on the MOSI dataset, using our Cross-modality Fusion Transformer to fuse language features and hyper-modality features is the most effective. On the CH-SIMS dataset, although TFN achieves better performance on the MAE metric, its Acc-5 is lower. Overall, using Transformer for feature fusion is effective.

Table 5: Effects of different hierarchical language guidance.

| $H_t^{(0)}$ | $H_t^{(1)}$ | $H_t^{(2)}$ | CH-SIMS | | MOSI | |
|---|---|---|---|---|---|---|
| | | | Acc-5 | MAE | Acc-7 | MAE |
| ✓ | | | 38.29 | 0.444 | 34.40 | 0.952 |
| | ✓ | | 43.54 | 0.412 | 47.38 | 0.704 |
| | | ✓ | 43.11 | 0.415 | 48.10 | 0.709 |
| ✓ | ✓ | | 45.51 | 0.417 | 46.36 | 0.736 |
| ✓ | | ✓ | 44.20 | 0.409 | 48.10 | 0.707 |
| | ✓ | ✓ | 43.76 | 0.416 | 47.81 | 0.729 |
| ✓ | ✓ | ✓ | **45.73** | **0.404** | **49.42** | **0.683** |

**Effects of Different Hierarchical Language Guidance.** To study the influence of different

[1] https://github.com/thuiar/MMSA/blob/master/results/result-stat.md

hierarchical text features, we report ablation results in Table 5. The results indicate that utilizing all hierarchical language scales ($H_t^{(0)}, H_t^{(1)}, H_t^{(2)}$) in the CAHL module yields the best performance, proving that multi-scale textual anchors are essential for guiding hyper-modality learning.

### 4.4 Visualization and Analysis

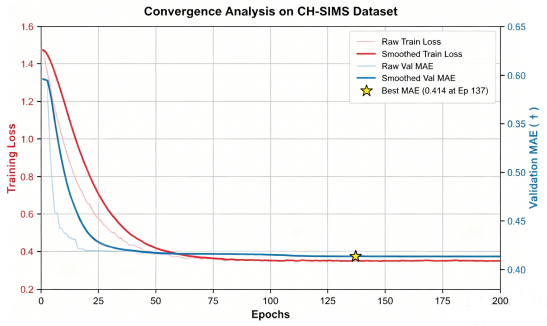



Figure 4: Convergence Curve of CHML on CH-SIMS.

**Convergence and Robustness Analysis.** Figure 4 tracks CHML's training loss and validation MAE on CH-SIMS across 200 epochs. The training loss descends continuously, while the validation MAE rapidly stabilizes to an optimal minimum of 0.414 at epoch 137. This healthy, divergence-free convergence demonstrates that explicitly gating modality conflicts (MCG) and adaptively disentangling features (CAMI) effectively prevent overfitting to noisy outliers, ensuring robust generalization.

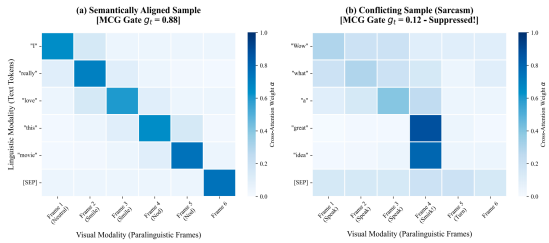



Figure 5: Visualization of attention dynamics in the CAHL module on CH-SIMS.

**Visualization of Conflict-Aware Attention Dynamics.** Figure 5 visualizes the cross-modal attention matrices and MCG activations (e.g., $g_t$) from the final CAHL layer on CH-SIMS. To illustrate our dynamic regulation, we contrast a semantically aligned sample with a conflicting (sarcastic) one. In Figure 5(a), when textual semantics align with paralinguistic cues, the MCG maintains a high activation ($g_t > 0.85$), allowing the hyper-modality to uniformly integrate synergistic audio-visual cues. Conversely, under severe semantic conflict (Figure 5(b)), the MCG detects the incongruity and drops sharply ($g_t < 0.2$). Simultaneously, the attention matrix becomes highly sparse and localized, focusing strictly on critical paralinguistic frames while ignoring deceptive text. This explicitly validates that CHML rejects blind fusion, rigorously shielding the global consensus from semantic pollution.

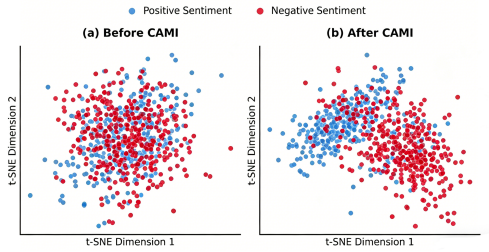



Figure 6: t-SNE visualization of feature distributions before and after disentanglement.

**Visualization of Feature Disentanglement.** To validate the CAMI module, we visualize feature distributions via t-SNE before and after Adaptive Residual Disentanglement. As shown in Figure 6, raw multimodal features are highly entangled, complicating sentiment classification near decision boundaries. Conversely, after CAMI processing, the pristine modality-specific features form distinct, sentiment-polarized clusters. This sharper boundary demonstrates that distilling unshared, modality-specific variations successfully uncovers true emotional states hidden beneath conflicting consensus.

## 5 Conclusion

In this paper, we propose Conflict-Aware Hyper-Modality Learning to tackle semantic pollution caused by inter-modal conflicts in Multimodal Sentiment Analysis. Specifically, a Modality Conflict Gate dynamically filters contradictory interactions to ensure a robust global consensus, while the CAMI module explicitly distills complementary paralinguistic cues under hierarchical linguistic guidance. Experiments on three benchmarks demonstrate that CHML achieves state-of-the-art performance, proving exceptionally effective in complex scenarios like sarcasm. Future work will extend this conflict-aware mechanism to multimodal Large Language Models.

## Limitations

Although the proposed CHML framework achieves highly competitive results, it still presents certain limitations. First, similar to other deep Transformer-based architectures, CHML is highly parameter-intensive and inherently data-hungry, requiring large-scale datasets for optimal training. However, current multimodal sentiment analysis datasets (e.g., CMU-MOSI and CH-SIMS) are generally limited in scale, which restricts the upper bound of the model's performance. For instance, compared to coarse-grained classification metrics like Acc-2 and Acc-7, fine-grained continuous regression metrics such as MAE and Corr impose much stricter demands on the volume of training data. Consequently, the margin of improvement achieved by CHML on these regression metrics is relatively constrained compared to its significant gains in classification accuracy. Second, our method currently focuses on text-audio-visual modalities, and due to the computational overhead of the cyclic feedback mechanism, it may require further structural optimizations for real-time inference scenarios.

## Ethics Statement

This research is conducted in strict accordance with the ACL Ethics Policy. The datasets utilized in this study (CH-SIMS, CMU-MOSI, and CMU-MOSEI) are publicly available and widely adopted in the academic community. We strictly adhere to their respective licenses, utilizing the data solely for academic research purposes. Although these multimodal datasets contain human facial and acoustic features, the data has been desensitized, and our framework does not attempt to infer any personally identifiable information (PII). Furthermore, the proposed method is intended to enhance human-computer interaction and affective computing; it is not designed for, nor should it be applied to, unethical surveillance or malicious psychological profiling. We declare that no unethical operations or privacy violations are involved in this work.

## References


Zhao, K., Zheng, M., Li, Q., & Liu, J. (2025). Multimodal Sentiment Analysis—A Comprehensive Survey From a Fusion Methods Perspective. IEEE Access, 13, 64556–64583.

Baltrušaitis, T., Ahuja, C., & Morency, L.P. (2018). Multimodal machine learning: A survey and taxonomy. IEEE Transactions on Pattern Analysis and Machine Intelligence, 41(2), 423–443.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. arXiv preprint arXiv:1706.03762.

Dhanith P R, J., Venkatraman, S., Sharma, V., & Malarvannan, S. (2024). Multimodal emotion recognition using audio-video transformer fusion with cross attention. arXiv preprint arXiv:2407.18552.

Wang, Y., et al. (2023). Causal Intervention for Multimodal Sentiment Analysis. In Proceedings of the 31st ACM International Conference on Multimedia (MM '23), 4210–4219.

Lin, J., et al. (2023). Hypergraph-enhanced Multimodal Learning for Sentiment Analysis. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP), 8192–8204.

Hazarika, D., Zimmermann, R., & Poria, S. (2020). MISA: modality-invariant and -specific representations for multimodal sentiment analysis. In Proceedings of the 28th ACM International Conference on Multimedia (MM '20), 1122–1131.

Zhang, H., Wang, Y., Yin, G., Liu, K., Liu, Y., & Yu, T. (2023). Learning language-guided adaptive hyper-modality representation for multimodal sentiment analysis. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, 756–767.

Wang, D., Liu, S., Wang, Q., Tian, Y., He, L., & Gao, X. (2022). Cross-modal enhancement network for multimodal sentiment analysis. IEEE Transactions on Multimedia, 25, 4909–4921.

Wang, X., Jiang, X., et al. (2024). Mitigating Modality Dominance in Multimodal Sentiment Analysis via Distribution Alignment. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 14230–14240.

Liu, Y., Ding, Y., et al. (2023). Sarcasm Detection via Multimodal Incongruity Modeling and Semantic Noise Reduction. In Proceedings of the 31st ACM International Conference on Multimedia (MM '23), 2104–2113.

Chen, Z., Shen, F., et al. (2024). Robust Multimodal Sentiment Analysis against Modality Conflict and Missing. IEEE Transactions on Affective Computing, 15(3), 812–825.

Zadeh, A., Chen, M., Poria, S., Cambria, E., & Morency, L.P. (2017). Tensor fusion network for multimodal sentiment analysis. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing.

Tsai, Y.H.H., Bai, S., Liang, P.P., Kolter, J.Z., Morency, L.P., & Salakhutdinov, R. (2019). Multimodal transformer for unaligned multimodal language sequences. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 6558–6569.

Wang, Y., et al. (2023). GCNet: Graph-based contrastive network for multimodal sentiment analysis. In Proceedings of the 31st ACM International Conference on Multimedia.

Yuan, H., et al. (2022). MM-GNN: A multimodal graph neural network for multimodal fusion and sentiment analysis. In Proceedings of the 2022 International Joint Conference on Neural Networks (IJCNN).

Li, Y., Wang, Y., & Cui, Z. (2023). Decoupled Multimodal Distilling for Emotion Recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 6631–6640.

Yang, J., Yu, Y., Niu, D., Guo, W., & Xu, Y. (2023). ConFEDE: Contrastive Feature Decomposition for Multimodal Sentiment Analysis. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics, 7617–7630.

Zhao, S., et al. (2025). SDRS: Sentiment-aware disentangled representation shifting for multimodal sentiment analysis. IEEE Transactions on Affective Computing, 16(3), 1802–1813.

Wang, H., et al. (2024). Multi-modal Relation Distillation for Unified 3D Representation Learning. arXiv preprint arXiv:2407.14007.

Li, M., et al. (2023). Contrastive Feature Decoupling for Multimodal Sarcasm Detection. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL), 1205–1218.

Chen, X., et al. (2024). Decoupling and Reorganizing: Disentangled Representation Learning for Multimodal Sentiment Analysis. IEEE Transactions on Multimedia, 26, 3105–3117.

Arevalo, J., Solorio, T., Montes-y-Gómez, M., & González, F.A. (2017). Gated multimodal units for information fusion. arXiv preprint arXiv:1702.01992.

Gao, Z., et al. (2024). Enhanced Experts with Uncertainty-Aware Routing for Multimodal Sentiment Analysis. In Proceedings of the 32nd ACM International Conference on Multimedia (MM '24), 9650–9659.

Gao, Z., et al. (2024). Embracing Unimodal Aleatoric Uncertainty for Robust Multimodal Fusion. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 26876–26885.

Liu, S., et al. (2024). Noise-Robust Modality Fusion for Multimodal Sentiment Analysis. IEEE Transactions on Affective Computing, 15(2), 482–495.

Wu, T., et al. (2025). Handling Modality Inconsistency in Multimodal Sentiment Analysis via Uncertainty-Aware Routing. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 33, 110–123.

Li, K., et al. (2024). Hierarchical text-guided refinement network for multimodal sentiment analysis. In Proceedings of the AAAI Conference on Artificial Intelligence.

Wang, P., Zhou, Q., Wu, Y., Chen, T., & Hu, J. (2025). DLF: Disentangled-language-focused multimodal sentiment analysis. In Proceedings of the AAAI Conference on Artificial Intelligence, 39(2), 1824–1832.

Sun, Z., et al. (2025). DEQA: Descriptions enhanced question-answering framework for multimodal aspect-based sentiment analysis. In Proceedings of the AAAI Conference on Artificial Intelligence.

Zhao, L., et al. (2024). Chain-of-Thought Reasoning for Multimodal Sentiment Analysis. In Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP), 9230–9244.

Li, M., Wang, Y., et al. (2024). Evaluating Large Vision-Language Models on Multimodal Sarcasm Detection. In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL).

Yang, H., Chen, X., et al. (2024). Towards Interpretable Multimodal Sentiment Analysis: A Paralinguistic Perspective. IEEE Transactions on Affective Computing.

Zhang, Z., Liu, S., et al. (2024). Multimodal Sarcasm Detection via Incongruity-Aware Graph Neural Networks. In Proceedings of the AAAI Conference on Artificial Intelligence.

Wu, T., Jiang, X., et al. (2024). Handling Modality Inconsistency in Multimodal Sentiment Analysis via Uncertainty-Aware Routing. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 32, 110–123.

Zhou, K., Hu, D., et al. (2024). Dynamic Modality Weighting for Robust Multimodal Sentiment Analysis. In Proceedings of the 32nd ACM International Conference on Multimedia (MM '24).

Wang, S., Mao, X., & Chen, W. (2020). 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, 3718–3727.

Zadeh, A., Liang, P.P., Poria, S., Vij, P., & Morency, L.P. (2018). Multimodal Language Analysis in the Wild: CMU-MOSI Dataset and Interpretable Dynamic Fusion Graph. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2236–2246.

Zadeh, A., Liang, P.P., Poria, S., Vij, P., & Morency, L.P. (2018). 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), 2236–2246.

Liu, Z., Shen, Y., Lakshminarasimhan, V.B., Liang, P.P., Zadeh, A., & Morency, L.P. (2018). 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), 1234–1244.

Singh, A., & Kose, N. (2020). Multimodal fusion module for unaligned multimodal sequences. arXiv preprint arXiv:2006.12701.

Lv, F., Chen, X., Huang, Y., Duan, L., & Lin, G. (2021). Progressive Modality Reinforcement for Human Multimodal Emotion Recognition from Unaligned Multimodal Sequences. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2554–2562.

Li, G., et al. (2022). MAG-BERT: Multimodal adaptive graph BERT for multimodal sentiment analysis. IEEE Transactions on Multimedia, 24, 3802–3814.

Guo, Z., et al. (2021). Self-MM: Self-supervised multimodal learning for multimodal sentiment analysis. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, 8650–8662.

Cai, Y., et al. (2022). MMIM: Multimodal memory interaction module for multimodal sentiment analysis. arXiv preprint arXiv:2204.03045.

Yang, D., Huang, S., Kuang, H., Du, Y., & Zhang, L. (2022). Disentangled Representation Learning for Multimodal Emotion Recognition. In Proceedings of the 30th ACM International Conference on Multimedia, 4548–4556.

Wang, X., et al. (2022). Cross-modal hierarchical fusion network for multimodal sentiment analysis. IEEE Access, 10, 128564–128575.

# A Hyper-parameters

In this section, we show the selection of some key hyper-parameters on the validation set of CH-SIMS.

## A.1 Overview

We used PyTorch to implement our method. The experiments were conducted on a high-performance Linux server, and the model training was accelerated using a single NVIDIA RTX A6000 (48GB) GPU with CUDA 12.8. The key parameters are shown in Table 6. We see that most hyper-parameters are the same across these datasets, demonstrating that CHML does not require complex hyper-parameters adjustment.

Table 6: Hyper-parameters of CHML we use on the different datasets.

| Hyper-parameter | MOSI | MOSEI | CH-SIMS |
|---|---|---|---|
| Modality Feature Length $T$ | 8 | 8 | 8 |
| Vector Dimension $d_h$ | 128 | 128 | 128 |
| Modality Embedding Depth | 1 | 1 | 1 |
| CAHL Depth | 3 | 3 | 3 |
| Fusion Transformer Depth | 2 | 4 | 4 |
| Batch Size | 64 | 64 | 64 |
| Initial Learning Rate | 1e-4 | 1e-4 | 1e-4 |
| Weight Decay | 1e-4 | 1e-4 | 1e-4 |
| Optimizer | AdamW | AdamW | AdamW |
| Epochs | 200 | 200 | 200 |
| Warmup Epochs | - | - | 10 |
| Random Seed | 7 | 0 | 8 |

## A.2 Effects of Length Settings of Modality Feature

In Figure 7, we show the effect of the sequence length $T$ of the token $H_m^{(0)}$ in modality embedding on the CH-SIMS dataset. It is observed that there are significant performance changes when the hyper-parameter is changed. And a similar phenomenon occurred on the MOSI and MOSEI datasets. Although the MAE is not the best when the Acc-5 is highest, e.g., $T$ is set to 32. Considering that the model computation rises when the $T$ increases, we set $T$ to 8, which is beneficial for CHML to obtain the best performance with a relatively lower computational cost.

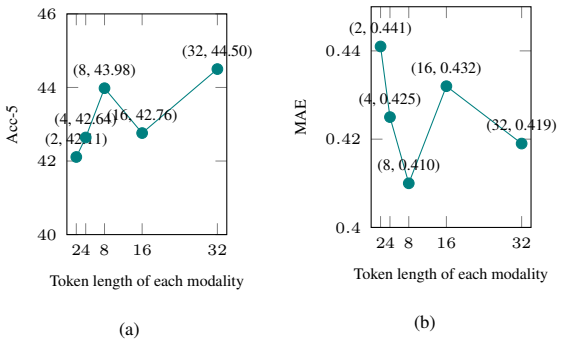



Figure 7: Effects of Token Length Settings in Modality Embedding. (a) Accuracy curve of different Token sequence lengths on the CH-SIMS; (b) MAE curve of different Token sequence lengths on the CH-SIMS.

## A.3 Effects of Depth Settings of CAHL

Figure 8 presents the effect of CAHL depth settings for MSA. Obviously, the CHML achieves the best performance on the two most difficult evaluation metrics, i.e., Acc-5 and MAE. Hence, in this study, we set the depth of CAHL to 3. Moreover, on the MOSI and MOSEI datasets, we set it to 3 as the similar phenomenon is also observed.

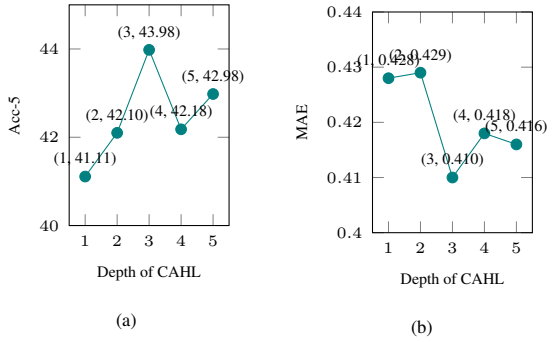
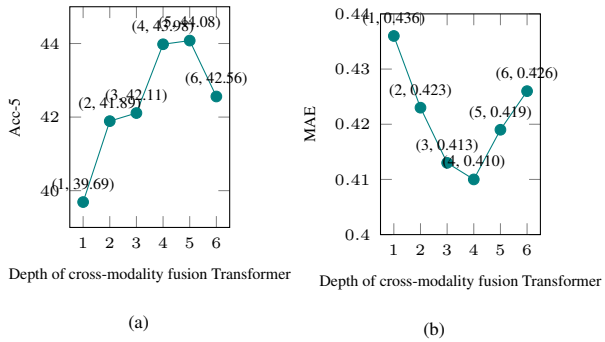



Figure 8: Effects of Depth Settings of CAHL. (a) Accuracy curve of different CAHL depth on the CH-SIMS; (b) MAE curve of different CAHL depth on the CH-SIMS.

## A.4 Effects of Depth Settings of Fusion Transformer

In Figure 9, we present the effects of depth settings of cross-modality fusion Transformer on CH-SIMS. We observed that the CHML can obtain the best result on Acc-5 and MAE when the depth is set to 5 and 3, respectively. However, to balance performance and model computation, we set the depth to 4 on the CH-SIMS. Following the similar rule, we set the depth of the cross-modality fusion Transformer to 2 and 4 on the MOSI and MOSEI, respectively.




(a) (b)

Figure 9: Effects of Depth Settings of Cross-modality Fusion Transformer. (a) Accuracy curve of different depth settings on the CH-SIMS; (b) MAE curve of different depth settings on the CH-SIMS.