CoMoR: A Conflict-Aware Modality Routing Framework for Multimodal Learning

Anonymous authors
Paper under double-blind review

Abstract

Multimodal learning combines complementary evidence from heterogeneous sources, but high confidence in each modality's prediction does not guarantee agreement across modalities. Therefore, a general challenge is to determine how modality-specific judgments should influence joint prediction. We propose CoMoR, a Conflict-Aware Modality Routing framework. The framework uses these judgments, their disagreement, and uncertainty to guide multimodal integration. CoMoR does not combine diagnostic predictions into a final answer. It uses them to adjust the audio-visual inputs, which a frozen large language model then processes together with the original text to predict sentiment. We evaluate CoMoR on four widely used English and Chinese sentiment benchmarks. Ablations on CMU-MOSI and CH-SIMS support the use of diagnostic information and sample-specific routing. Among the methods compared, CoMoR obtains the lowest mean absolute error on all four datasets and leads on most reported metrics, with competitive results on the remaining ones.

1 Introduction

Multimodal learning combines heterogeneous sources to draw on complementary task evidence (Baltrušaitis et al., 2019). For the same sample, different modalities can yield confident but conflicting predictions (Bezirganyan et al., 2025). We study how to incorporate such evidence when adapting frozen large language models (LLMs). Audio-visual information is introduced through learned continuous inputs (Tsimpoukelli et al., 2021; Yang et al., 2025).

Existing fusion methods adjust modality contributions according to uncertainty, data quality, or conflict (Han et al., 2023; Zhang et al., 2023b; Bezirganyan et al., 2025; Wang et al., 2026; Zou & Wei, 2026). Methods for multimodal language-model adaptation construct continuous inputs and prompts to introduce non-textual evidence (Yang et al., 2025; Zhao et al., 2026; Lin et al., 2026). Related approaches also adjust prompts within frozen encoders (Kou et al., 2026) or reweight decoding branches (Deng et al., 2026). For input-side adaptation, an important question is what task-level evidence should guide the adjustment of modality contributions before joint prediction.

Conflicting predictions alone are not enough to decide which modality is right or which one to downweight. Modalities may carry shared, unique, or synergistic task information (Dufumier et al., 2025), so different modalities may make different predictions because they provide different cues. Conflict refers to disagreement among modality-specific predictions for the same target, which can arise even when the inputs are neither contradictory nor corrupted. Predictive entropy reflects how concentrated each modality's predictive distribution is, whereas disagreement captures differences between these distributions. Neither alone establishes which modality is correct. The challenge is therefore to use modality-specific predictions, together with their uncertainty and disagreement, to guide input adaptation while preserving useful evidence for joint prediction.

To tackle those challenges, we propose CoMoR, a Conflict-Aware Modality Routing framework that uses modality-specific diagnostic distributions to guide audio-visual input scaling. The full diagnostic distributions retain sentiment-level predictions beyond a scalar confidence score. They guide audio-visual pseudo-token scaling, while a frozen LLM predicts sentiment from the scaled blocks and original text.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

We instantiate CoMoR in multimodal sentiment analysis (MSA), where verbal and nonverbal cues can support different sentiment judgments (Yu et al., 2020; Liu et al., 2022). Shared Sentiment Diagnosis (SSD) produces a probability distribution over common ordered sentiment anchors for each modality. We use normalized Wasserstein-1 distance to quantify pairwise disagreement on the shared sentiment scale, accounting for the distances between sentiment anchors. For routing, we feed the diagnostic distributions, their pairwise disagreement, and predictive entropy into Conflict- and Uncertainty-Aware Token Routing (CUTR). It learns a scaling factor for each audio-visual pseudo-token block. Our contributions are as follows:

- We present CoMoR, a framework that uses modality-specific diagnostic predictions to effectively scale audio-visual pseudo-token blocks for frozen LLMs. Instead of fusing these distributions to produce a final answer, CoMoR uses them to adjust the pseudo-token blocks, enabling robust joint prediction alongside the original text sequence.
- We develop SSD and CUTR. SSD employs a shared frozen LLM alongside a shared trainable diagnostic head to produce specific diagnostic distributions aligned to a unified sentiment scale. CUTR uses these distributions, their normalized Wasserstein-1 distances, and predictive entropy to learn sample-specific scaling weights for audio-visual pseudo-token blocks. Importantly, higher disagreement does not automatically correspond to a lower scaling factor.
- We evaluate CoMoR across four sentiment benchmarks covering both English and Chinese. CoMoR surpasses baseline methods on the majority of metrics and maintains strong performance on the remaining ones. These results validate CoMoR's effectiveness for MSA across the four MSA benchmarks.

2 Related Work

Multimodal representations and interactions. DecAlign separates shared and modality-specific representations through hierarchical alignment (Qian et al., 2026), while ITHP organizes multimodal information through hierarchical information bottlenecks (Xiao et al., 2024). CoMM examines shared, unique, and synergistic information in contrastive learning (Dufumier et al., 2025). Liang et al. (2024) link multimodal interactions to disagreement between optimal unimodal classifiers in semi-supervised learning. With suboptimal classifiers, disagreement may also reflect prediction errors. TCMax addresses modality competition during joint training through total correlation maximization (Yu et al., 2026). CoMoR uses sample-level predictive disagreement as one signal for scaling audio-visual pseudo-token blocks before joint prediction with a frozen LLM. We do not estimate shared, unique, or synergistic information. Nor do we use a dedicated loss to balance learning across modalities.

Uncertainty and conflict-aware fusion. Trusted Multi-View Classification combines Dirichlet evidence using Dempster–Shafer theory (Han et al., 2023), while QMF uses uncertainty in dynamic fusion (Zhang et al., 2023b) and Discounted Belief Fusion adjusts conflicting evidence through discounting (Bezirganyan et al., 2025). These methods differ in how they assess the modalities and which part of the fusion process they modify. CACR measures conflict in a shared feature space to weight cross-reconstruction (Wang et al., 2026). CoRiM optimizes decision-fusion weights through a risk objective over predictive distributions (Zou & Wei, 2026), whereas CAD assesses predictive conflict and its actionability to redistribute joint and unimodal decoding weights without training (Deng et al., 2026). CoMoR routes evidence before joint prediction: it scales audio-visual pseudo-token blocks instead of averaging diagnostic predictions. The frozen LLM can therefore use information in the scaled blocks and original text beyond what is summarized by the diagnostics.

Multimodal adaptation of language models. MAG uses audio and visual cues to adjust text representations while fine-tuning BERT and XLNet (Rahman et al., 2020). Frozen uses learned visual embeddings to condition a fixed language model (Tsimpoukelli et al., 2021). MSE-Adapter converts audio and visual features into pseudo-tokens for sentiment prediction (Yang et al., 2025). In Prototype-as-Prompt, sentiment prototypes provide the basis for constructing prompts (Zhao et al., 2026). MGSI combines audio-visual features learned at short, medium, and long time scales (Lin et al., 2026). VG-TPT uses text and visual features to route prompts within a frozen text encoder (Kou et al., 2026). CoMoR uses a shared frozen LLM and a shared trainable diagnostic

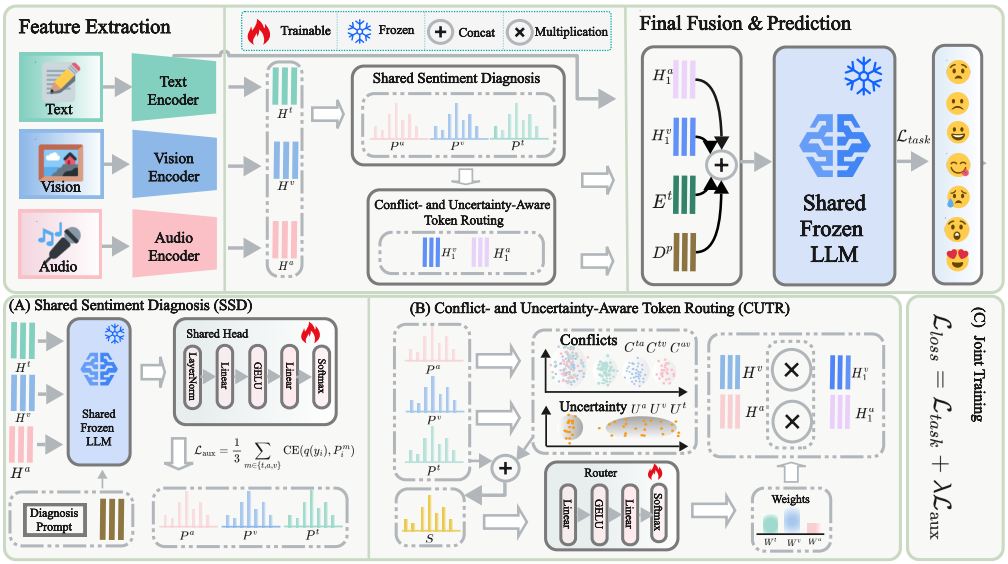

Figure 1: **Overall architecture of the proposed CoMoR framework.** (A) SSD uses a shared frozen LLM and a shared trainable diagnostic head. Then it produces modality-specific distributions over sentiment anchors. (B) CUTR uses these distributions, their normalized Wasserstein-1 distances, and predictive entropy to scale audio-visual pseudo-token blocks. The same frozen LLM then predicts sentiment from the scaled blocks and the original text sequence. (C) The trainable modules are jointly optimized with the task loss and auxiliary diagnostic loss.

head to obtain modality-specific diagnostic distributions. The router uses these distributions, their pairwise disagreement, and predictive entropy to decide how much to scale the audio-visual pseudo-token blocks before the final prediction. The original text sequence is passed directly to the LLM without router-based scaling.

3 METHOD

Overall architecture and problem formulation. As shown in Figure 1, CoMoR first summarizes the text, audio, and visual inputs and projects the resulting features into modality-specific pseudo-token blocks (Section 3.1). Then SSD produces diagnostic distributions using a shared frozen LLM and a shared trainable diagnostic head (Section 3.2). CUTR uses the diagnostic information to scale the audio-visual pseudo-token blocks (Section 3.3). The same frozen LLM makes the final prediction from the pseudo-token blocks and original text.

Each sample in a multimodal dataset contains text, audio, and visual inputs, indexed by $\mathcal{M} = \{t, a, v\}$. Given a text sequence x^t and audio and visual features x^a and x^v, the task is to predict a sentiment score $\hat{y}$. F_ϕ denotes the frozen LLM with fixed parameter ϕ during training.

3.1 FEATURE EXTRACTION

We convert each modality's features into pseudo-token blocks. The blocks serve as inputs to the frozen LLM. Separate LSTM encoders deal with the audio and visual sequences (Hochreiter & Schmidhuber, 1997). Their final hidden states are linearly projected to $f^m \in \mathbb{R}^{d_f}$ for $m \in \{a, v\}$, where d_f is the output feature dimension. For the text modality, we apply attention pooling to token embeddings from the frozen layer E_ϕ. A linear projection followed by GELU produces $f^t \in \mathbb{R}^{d_f}$. We retain the original ordered text embeddings E^t for joint prediction. Each modality has its own projector A_m. The projectors follow the architecture of MSE-Adapter (Yang et al., 2025). Each projector converts f^m into pseudo-tokens, to which we add a learned modality embedding to obtain $H^m \in \mathbb{R}^{N_p \times d}$. Here, N_p is the number of pseudo-tokens and d is the LLM embedding dimension. Appendix A provides the implementation details.

3.2 SHARED SENTIMENT DIAGNOSIS

Feature differences do not necessarily imply disagreement about sentiment. SSD places modality-specific sentiment judgments on a common scale for comparison. Each modality is processed independently using the same frozen LLM and diagnostic prompt. A shared trainable diagnostic head produces a distribution over K sentiment anchors. The anchors satisfy $a_1 < \cdots < a_K$ and span the target range. Sharing the head applies the same mapping from LLM hidden states to sentiment anchors across modalities. The input includes a pseudo-token block H and prompt embeddings D. We write it as $\mathcal{I}(H; D) = [E_{\mathrm{BOS}}; E_{\mathrm{open}}; H; E_{\mathrm{close}}; D]$. Here, semicolons inside the brackets denote concatenation. E_{BOS} marks the sequence beginning, while E_{open} and E_{close} delimit the pseudo-token block. We obtain the hidden state h^m:

$$h^m = F_\phi(\mathcal{I}(H^m; D_{\mathrm{diag}}))_{\mathrm{last}} \in \mathbb{R}^d, \tag{1}$$

where D_{diag} denotes the shared diagnostic prompt and $(\cdot)_{\mathrm{last}}$ selects the hidden state at the last valid prompt position. The shared diagnostic head maps h^m to the sentiment-anchor logits s^m:

$$s^m = \mathrm{Linear}_2(\mathrm{Dropout}(\mathrm{GELU}(\mathrm{Linear}_1(\mathrm{LN}(h^m))))) \,. \tag{2}$$

Here, LN denotes layer normalization. We finally obtain these distributions by applying softmax:

$$P^m = \mathrm{softmax}_K(s^m) \in \Delta^{K-1}, \tag{3}$$

where Δ^{K-1} denotes the probability simplex.

3.3 CONFLICT- AND UNCERTAINTY-AWARE TOKEN ROUTING

Disagreement alone does not identify which modality to downweight. A concentrated diagnostic distribution can also be incorrect. CUTR therefore combines the diagnostic distributions with their disagreement and uncertainty to learn sample-specific input scaling. The full distributions describe what each modality predicts. Disagreement and entropy are supplied explicitly as summaries of these distributions, not as additional evidence. Providing these summaries explicitly means the router need not derive them from the full distributions.

Predictive disagreement and entropy. Sentiment levels are ordered, and disagreement should reflect how far apart they are. We thus use normalized Wasserstein-1 distance (Peyré & Cuturi, 2019):

$$C^{mn} = \frac{1}{a_K - a_1} \sum_{k=1}^{K-1} (a_{k+1} - a_k) \left| \sum_{\ell=1}^{k} (P_\ell^m - P_\ell^n) \right|, \quad (m, n) \in \{(t, a), (t, v), (a, v)\}, \tag{4}$$

where the anchor gaps account for sentiment separation, and normalization gives $C^{mn} \in [0, 1]$. Entropy reflects how concentrated the diagnostic distributions are. Together with these distributions and their disagreement, it guides sample-specific routing. For each modality, we compute normalized predictive entropy:

$$U^m = -\frac{1}{\log K} \sum_{k=1}^{K} P_k^m \log P_k^m. \tag{5}$$

Here, $U^m \in [0, 1]$ is lower for more concentrated distributions. We concatenate the diagnostic distributions, pairwise disagreement, predictive entropy, and modality-availability indicators to form the router input:

$$S = [P^t; P^a; P^v; C^{ta}; C^{tv}; C^{av}; U^t; U^a; U^v; b^t; b^a; b^v] \in \mathbb{R}^{3K+9}. \tag{6}$$

Here, b^m indicates whether modality m is included. The following routing equations describe complete inputs, where $b^t = b^a = b^v = 1$. Semicolons denote concatenation along the coordinate axis. Appendix D specifies the missing-modality placeholders and masking rules.

Sample-specific block routing. We use SSD to obtain a separate sentiment judgment for each modality. Combining these judgments directly would make the final score depend on the diagnostic distributions alone. We instead retain the original text and audio-visual pseudo-token blocks for the frozen LLM to process jointly. The diagnostic information guides the scaling of the audio-visual blocks before this joint prediction. The contribution of a modality can vary across samples. CUTR therefore predicts sample-specific routing weights from S, rather than assigning fixed weights. The router maps S to routing logits:

$$u = \mathrm{Linear}_3(\mathrm{Dropout}(\mathrm{GELU}(\mathrm{Linear}_4(\mathrm{sg}(S))))) . \tag{7}$$

Here, sg denotes stop-gradient, and u^m is the routing logit for modality m. Both linear layers are trainable. A softmax converts the logits into routing weights:

$$W^m = \frac{\exp(u^m)}{\sum_{n\in\mathcal{M}} \exp(u^n)}. \tag{8}$$

Here, W^m is the weight assigned to modality m. The routing weights scale the audio-visual blocks as follows:

$$H_1^a = 3W^a H^a, \qquad H_1^v = 3W^v H^v. \tag{9}$$

Each scaling factor applies to the entire block, including its modality embedding. The factor 3 gives unit scaling under uniform routing. The original text sequence is not scaled. Its routing weight W^t affects the audio-visual scaling factors through their shared normalization: the two factors sum to $3(1 - W^t)$. Higher disagreement or entropy does not necessarily lead to a smaller scaling factor.

3.4 Final Fusion and Prediction

The same frozen LLM makes the final prediction from the scaled audio-visual pseudo-token blocks and the original text. We construct the joint input as:

$$I_{\text{joint}} = [E_{\text{BOS}}; E_{\text{open}}; H_1^a; H_1^v; E_{\text{close}}; E^t; D^p], \tag{10}$$

where D^p contains the embeddings of a task prompt requesting a numerical sentiment score. Text pseudo-tokens H^t are used only for diagnosis; joint prediction uses the original text embeddings E^t. All textual embeddings come from E_ϕ. Padding positions are masked without changing the order of valid tokens. At inference, diagnosis and routing are computed once per sample. The frozen LLM generates a numerical response using greedy decoding. We parse this response as the sentiment score $\hat{y}$. Appendix F specifies the prompts, response format, decoding limits, and handling of invalid outputs.

3.5 Joint Training

CoMoR learns both modality-specific diagnosis and joint prediction. The diagnostic distributions need supervision on the common sentiment scale, while the final response must match the target score. We therefore combine an auxiliary diagnostic loss with a generative prediction loss. Both objectives use the same sample-level sentiment label, without requiring modality-specific annotations.

Diagnostic supervision. The sentiment labels are continuous, but SSD predicts probabilities over discrete anchors. Assigning each label to a single anchor would discard its position between neighboring sentiment levels. We therefore distribute the target mass between the two adjacent anchors according to their distances from the label.

Let $\bar{y} = \min\{a_K, \max\{a_1, y\}\}$ denote the target clipped to the anchor range. For $\bar{y} \in [a_k, a_{k+1}]$, we define:

$$q_k(y) = \frac{a_{k+1} - \bar{y}}{a_{k+1} - a_k}, \qquad q_{k+1}(y) = \frac{\bar{y} - a_k}{a_{k+1} - a_k}. \tag{11}$$

All other entries of $q(y) \in \Delta^{K-1}$ are zero. At either endpoint, the corresponding anchor receives unit mass. The resulting target preserves the clipped score in expectation: $\sum_k a_k q_k(y) = \bar{y}$. We supervise the diagnostic distributions with cross-entropy:

$$\mathcal{L}_{\text{aux}} = \frac{1}{3} \sum_{m\in\mathcal{M}} \mathrm{CE}(q(y_i), P_i^m) . \tag{12}$$

Here, CE denotes cross-entropy. For each sample i, the loss averages over the three modalities, using the same target distribution for each modality. For prediction, let Y denote the token sequence of the numerical target $\bar{y}$, using the response format in Appendix F. We use response-token cross-entropy:

$$\mathcal{L}_{\text{task}} = \text{CE}_{\text{resp}}(Y \mid I_{\text{joint}}) . \tag{13}$$

Here, CE_{resp} denotes the cross-entropy under the frozen LLM, and each response token is predicted using the preceding target tokens. The overall training objective is:

$$\mathcal{L}_{\text{loss}} = \mathcal{L}_{\text{task}} + \lambda \mathcal{L}_{\text{aux}}, \tag{14}$$

where λ controls the contribution of diagnostic supervision.

Gradient flow. The LLM, vocabulary embeddings, and vocabulary output head remain frozen. Their parameters are not updated, but gradients still pass through the frozen LLM to its trainable inputs. We detach S before routing. This blocks task-loss gradients through the diagnostic descriptor, while allowing the router itself to learn from $\mathcal{L}_{\text{task}}$. The text diagnostic modules and shared diagnostic head are therefore trained only by $\mathcal{L}_{\text{aux}}$. The audio-visual encoders, projectors, and modality embeddings participate in both diagnosis and joint prediction. They receive auxiliary-loss gradients through the diagnostic paths and task-loss gradients through the scaled pseudo-token blocks. We detach S, but both losses still update the shared audio-visual parameters.

4 Experiments

4.1 Experimental Setup

Datasets and metrics. To comprehensively evaluate the performance of our CoMoR framework, we conduct experiments on four benchmark datasets widely used in MSA, namely two English datasets CMU-MOSI and CMU-MOSEI (Zadeh et al., 2016; Bagher Zadeh et al., 2018) and two Chinese datasets CH-SIMS and CH-SIMS v2 (Yu et al., 2020; Liu et al., 2022). Following previous work (Lin et al., 2026; Yang et al., 2025), we evaluate MSA performance from two aspects: classification and regression. The classification metrics include Acc-2 and F1-score, while the regression metrics include MAE and Corr. The detailed dataset splits, metrics, and calculation details are provided in Appendix G.2.

Implementation details. We use ChatGLM3-6B (Team GLM et al., 2024) as the backbone and keep its pre-trained parameters frozen throughout training. Only the newly added modules are optimized. We train the model with AdamW using an effective batch size of 16 for 40 epochs. The learning rate is linearly warmed up during the first 10% of the optimization steps and then decayed following a cosine schedule. We run CoMoR with three random seeds, 1111, 1113, and 1115, on NVIDIA RTX 3090 and RTX 4090 GPUs, and we report the mean across runs, while the standard deviations, dataset-specific learning rates, and other hyperparameters are reported in Appendix G.1 and Appendix H.1 due to space limitations.

Compared methods. We compare CoMoR with representative multimodal fusion and representation learning methods, including TFN (Zadeh et al., 2017), MulT (Tsai et al., 2019), MISA (Hazarika et al., 2020), Self-MM (Yu et al., 2021), ALMT (Zhang et al., 2023a), CLGSI (Yang et al., 2024), DLF (Wang et al., 2025), MDF (Jian et al., 2026) and DBR (Meng et al., 2026), as well as MISR (Zhu et al., 2025) and CACR (Wang et al., 2026). We also compare with MSE-Adapter (Yang et al., 2025), which uses audio-visual pseudo-tokens with the same frozen ChatGLM3-6B backbone.

4.2 Main Results

Table 1 compares CoMoR with existing methods on four English and Chinese sentiment benchmarks. CoMoR achieves the best results on 13 of the 16 metrics among the compared methods and stays competitive on the remaining metrics. CoMoR ranks first on all four metrics for CMU-MOSI and CH-SIMS v2. It performs well in both sentiment classification and score prediction. On CH-SIMS, CoMoR achieves the highest Acc-2 of 81.98%. It also obtains the lowest MAE and highest correlation on all four datasets. The comparison further evaluates CoMoR and MSE-Adapter with the same frozen LLM ChatGLM3-6B. CoMoR improves all metrics on both English and Chinese

Table 1: Comparison results across four benchmarks. Best results are highlighted in **bold**, and second-best results are underlined. * denotes baseline results obtained from the open-source MMSA benchmark (THUIAR, 2022). † denotes that the models use ChatGLM3-6B as the backbone. ‡ denotes results reproduced using the official source code. The results of other methods are directly taken from their original papers.

Models	CMU-MOSI				CMU-MOSEI				CH-SIMS				CH-SIMS v2			
	MAE↓	Corr↑	Acc-2↑	F1↑	MAE↓	Corr↑	Acc-2↑	F1↑	MAE↓	Corr↑	Acc-2↑	F1↑	MAE↓	Corr↑	Acc-2↑	F1↑
TFN* (2017)	0.947	0.673	79.08	79.11	0.573	0.714	81.89	81.74	0.432	0.591	78.38	78.62	0.303	0.707	80.14	80.14
MulT* (2019)	0.880	0.702	80.98	80.95	0.559	0.733	84.63	84.52	0.453	0.564	78.56	79.66	0.291	0.738	80.68	80.73
MISA* (2020)	0.777	0.778	83.54	83.58	0.558	0.752	84.67	84.66	0.447	0.563	76.50	76.60	0.342	0.671	78.20	78.30
Self-MM* (2021)	0.708	0.796	85.46	85.43	0.531	0.765	85.15	84.90	0.425	0.595	80.04	80.44	0.311	0.695	79.69	79.76
ALMT (2023)	0.683	0.805	86.43	86.47	0.526	0.779	86.79	86.86	0.404	0.619	81.19	81.57	–	–	–	–
CLGSI (2024)	0.703	0.790	86.43	86.25	0.532	0.763	86.32	86.18	0.408	0.634	81.18	80.59	–	–	–	–
DLF (2025)	0.722	0.781	85.06	85.04	0.536	0.764	85.42	85.27	–	–	–	–	–	–	–	–
MDF (2026)	0.692	0.810	85.82	85.90	0.525	0.788	86.61	86.41	–	–	–	–	–	–	–	–
DBR (2026)	0.681	0.811	86.87	86.83	0.526	0.788	86.73	86.78	–	–	–	–	–	–	–	–
MISR (2025)	0.671	0.819	88.11	88.15	0.513	0.789	**87.51**	**87.55**	0.425	0.597	81.53	**81.91**	–	–	–	–
CACR (2026)	0.710	–	87.04	86.90	0.532	–	85.37	85.36	0.423	0.596	81.62	80.99	–	–	–	–
MSE-Adapter†‡	0.672	0.834	88.38	88.30	0.516	0.786	86.61	86.31	0.395	0.626	79.47	77.39	0.306	0.705	81.43	81.30
CoMoR†	**0.581**	**0.864**	**89.28**	**89.21**	**0.493**	**0.796**	86.88	86.72	**0.349**	**0.684**	**81.98**	80.82	**0.271**	**0.752**	**83.88**	**83.76**

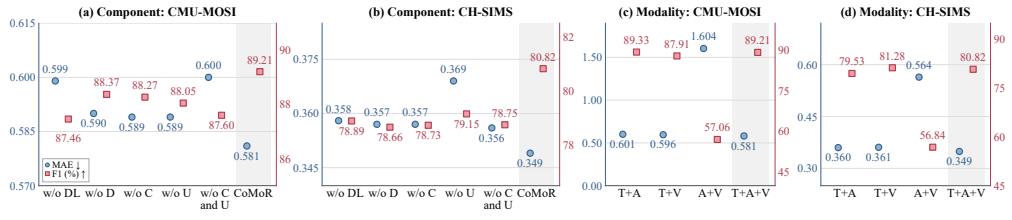

Figure 2: Component and modality ablation results. (a)–(b) Component ablations on CMU-MOSI and CH-SIMS. (c)–(d) Modality ablations on CMU-MOSI and CH-SIMS. DL denotes diagnostic LLM. D denotes diagnostic distribution inputs. C and U denote conflict inputs and entropy inputs.

datasets. On CMU-MOSI, Acc-2 and F1 increase by 0.90 and 0.91 percentage points, respectively. The improvements are larger on CH-SIMS and CH-SIMS v2, where F1 increases by 3.43 and 2.46 percentage points. These results support the general effectiveness of CoMoR for multimodal sentiment prediction without changing the LLM backbone. They also demonstrate CoMoR's overall strategy of using modality-specific diagnosis to guide audio-visual pseudo-token scaling before joint prediction by a frozen LLM.

4.3 Ablation Studies

Ablation definitions and settings are given in Appendix G.4. Key standard deviations of ablation results are shown in Appendix H.2. We also evaluate random token/frame missingness from 0% to 90% at test time, with settings in Appendix G.5 and results in Appendix H.3.

Component ablation. As shown in Figure 2 (a)–(b), the full model achieves the lowest MAE and the highest F1 on both datasets. In this ablation, we use mean-pooled pseudo-token blocks in place of the diagnostic LLM. This change reduces F1 by 1.75 percentage points on CMU-MOSI and 1.93 points on CH-SIMS. This supports the use of LLM-based diagnosis over direct adapter-based diagnosis for guiding subsequent input scaling. Dropping either the explicit conflict (C) or entropy (U) input degrades both metrics, even though the diagnostic distributions are still available to the router. The results suggest that using explicit summaries of cross-modal disagreement and within-modality uncertainty works better than using the distributions alone. Removing both summaries reduces F1 on CMU-MOSI from 89.21% to 87.60%, a drop of 1.61 percentage points, while removing entropy yields the largest MAE increase on CH-SIMS, from 0.349 to 0.369. Removing diagnostic distribution inputs (w/o D) worsens MAE and F1 on both datasets, supporting their use alongside conflict and entropy.

Table 2: Frozen LLM backbone comparison. Best and second-best values are bold and underlined. All other experimental settings remain unchanged except for the LLM backbone.

LLM	CMU-MOSI				CH-SIMS			
	MAE ↓	Corr ↑	Acc-2 ↑	F1 ↑	MAE ↓	Corr ↑	Acc-2 ↑	F1 ↑
Qwen-1.8B	0.874	0.682	81.50	81.15	<u>0.471</u>	0.458	<u>74.69</u>	72.26
Llama3.2-3B	<u>0.638</u>	<u>0.833</u>	<u>88.06</u>	<u>88.03</u>	0.475	<u>0.467</u>	74.25	<u>72.77</u>
Llama2-7B	0.746	0.767	84.25	84.00	0.489	0.417	72.94	68.43
ChatGLM3-6B	**0.581**	**0.864**	**89.28**	**89.21**	**0.349**	**0.684**	**81.98**	**80.82**

Modality ablation. Figure 2(c)–(d) shows the results for four trained modality configurations. The final prediction path in T-only has no trainable parameters and cannot be trained for the task. Its responses are often unparseable, and we omit it from the comparison. Both $T + A$ and $T + V$ have lower MAE and higher F1 than $A+V$ on both datasets, showing the benefit of retaining text for joint prediction. The full model has the lowest MAE: 0.581 on CMU-MOSI and 0.349 on CH-SIMS. On CMU-MOSI, $T + A$ has the highest F1 at 89.33, just 0.12 percentage points above the full model's 89.21. On CH-SIMS, $T+V$ reaches an F1 of 81.28, which is 0.46 percentage points higher than the full model. Text with audio or vision may be enough for binary classification. Even though binary F1 stays the same or decreases, the third modality may improve sentiment intensity prediction.

Diagnostic and routing choices. Table 3 compares diagnostic-head sharing, conflict measures, and routing strategies. Replacing the shared diagnostic head with independent heads lowers F1 by 2.56 percentage points on CMU-MOSI and 2.64 on CH-SIMS. Replacing only the three normalized Wasserstein-1 (W_1) conflict inputs with normalized cosine dissimilarities between diagnostic hidden vectors worsens both MAE and F1 on both datasets. Using normalized Jensen–Shannon (JS) divergence (Lin, 1991) instead also yields higher MAE on both datasets, although JS achieves higher F1 on CMU-MOSI (89.65% vs. 89.21%). For routing, the full dynamic router achieves lower MAE and higher F1 than both uniform and learned global gates on the two datasets, supporting sample-specific weighting. Feature-driven routing instead generates sample-specific weights from projected modality summaries using the same routing MLP architecture. Compared with this variant, CoMoR achieves lower MAE on both datasets (0.581 vs. 0.597 on CMU-MOSI and 0.349 vs. 0.354 on CH-SIMS) and higher F1 on CMU-MOSI (89.21% vs. 87.35%). On CH-SIMS, however, feature-driven routing achieves an F1 of 81.39%, compared with 80.82% for CoMoR.

Table 3: Comparison of diagnostic and routing design choices. Best results are shown in bold.

Variant	CMU-MOSI		CH-SIMS	
	MAE ↓	F1 ↑	MAE ↓	F1 ↑
CoMoR (full)	**0.581**	89.21	**0.349**	80.82
Independent heads	0.605	86.65	0.364	78.18
Hidden cosine conflict	0.595	88.16	0.360	78.33
JS divergence	0.591	**89.65**	0.358	80.51
Uniform gates	0.600	88.52	0.363	76.46
Learned global gates	0.600	87.25	0.377	79.33
Feature-driven routing	0.597	87.35	0.354	**81.39**

4.4 Further Analysis

Performance across LLM backbones. We conduct our experiments on four frozen LLMs, namely Qwen-1.8B (Bai et al., 2023), Llama3.2-3B (Grattafiori et al., 2024), Llama2-7B (Touvron et al., 2023) and ChatGLM3-6B. ChatGLM3-6B leads on all eight reported metrics in Table 2. Llama3.2-3B is the strongest substitute LLM with an F1 of 88.03% on CMU-MOSI. It is slightly lower than ChatGLM3-6B's 89.21%. On CH-SIMS, ChatGLM3-6B achieves 80.82% F1, while the next-best backbone only reaches 72.77%. This 8.05-point gap is larger than the 1.18-point gap on CMU-MOSI. Size does not always play a decisive role. Other differences like architecture and pre-training may have an effect on the results. For example, Llama3.2-3B outperforms the larger Llama2-7B on both datasets. These results show that CoMoR can be used with different frozen LLMs and support ChatGLM3-6B as the backbone for our main English and Chinese experiments.

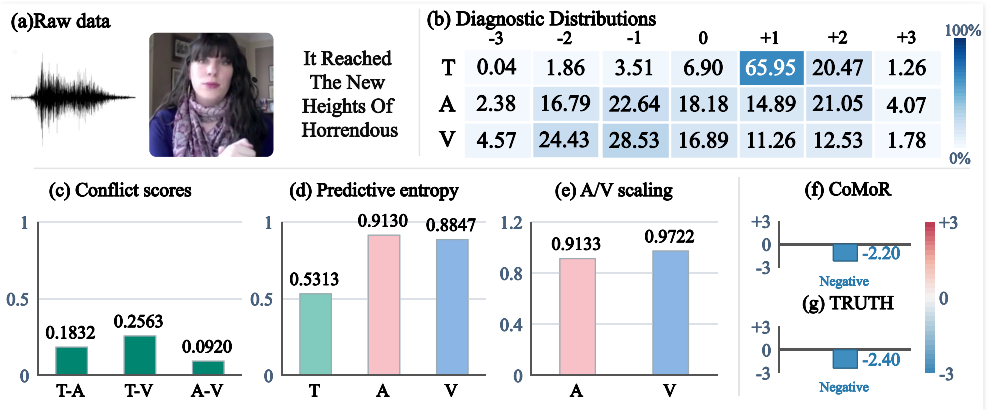

Figure 3: Case study on CMU-MOSI showing the raw data, diagnostic distributions, conflict scores, predictive entropy, audio-visual scaling factors, and the final prediction alongside the ground truth.

Case study. As shown in Figure 3, the text diagnosis assigns a 65.95% probability to the +1 anchor for the remark (It reached the new heights of horrendous), even though the remark is clearly negative and the visual diagnosis favors negative sentiment. Text and vision therefore show the largest pairwise conflict in this example, with a score of 0.2563. CUTR scales the audio and visual inputs by 0.9133 and 0.9722, respectively, reducing the audio input more than the visual input. Despite the misleading text diagnosis, the full model predicts -2.20, close to the ground-truth score of -2.40, with an absolute error of 0.20. Even though the text distribution is fairly concentrated, the final joint prediction does not simply follow the text diagnosis.

Parameter analysis. Figure 4 compares CoMoR with MSE-Adapter on CMU-MOSI. Both of them use the same frozen ChatGLM3-6B on NVIDIA RTX 3090. Trainable parameters increase from 2.61M to 6.72M, but still account for only about 0.11% of CoMoR's total parameters. The diagnostic stage reuses the frozen LLM. Peak training memory increases by only 0.17 GB, from 12.21 to 12.38 GB. The additional computation has a larger effect on runtime. The model training takes 217.66 seconds per epoch, approximately 1.95 times the 111.45 seconds required by MSE-Adapter. Inference latency increases by 21.8% and reaches 129.39 milliseconds per sample. CoMoR gives better prediction results than MSE-Adapter. Training is restricted to a small fraction of its parameters, with only minimal extra memory. However, training and inference take longer.

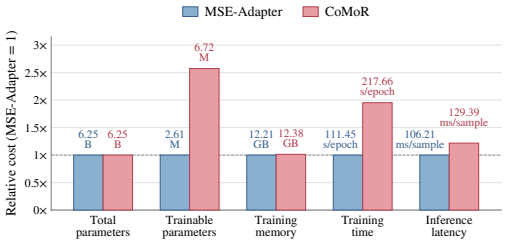

Figure 4: Computational cost of different methods on CMU-MOSI with ChatGLM3-6B. Training memory denotes peak GPU memory usage. Training time is measured per epoch, and inference latency is measured per sample.

5 Conclusion

We presented CoMoR, a framework that uses modality-specific diagnostic distributions to guide audio-visual input scaling for frozen LLMs. SSD produces these distributions over shared ordered sentiment anchors using a shared frozen LLM and a shared trainable diagnostic head. CUTR uses the diagnostic distributions, their normalized W1 distances, and predictive entropy to learn sample-specific scaling factors for audio-visual pseudo-token blocks. The same frozen LLM then makes the final prediction from the scaled blocks and the original text sequence. Among the methods we compare, CoMoR has the lowest MAE on all four English and Chinese sentiment benchmarks and leads on most metrics. Ablation studies support the effectiveness of the shared diagnostic head and sample-specific routing.

AI USE STATEMENT

We used ChatGPT to help revise the manuscript, find related work, design experiments, summarize results, and discuss the feasibility of the proposed method. It also helped write and modify small pieces of code. ChatGPT also drafted portions of the main body and appendix, including mathematical exposition and formulas, as well as descriptions of experimental settings and results. For the mathematical content, we used it to revise and check our existing derivations of the expectation property of the soft ordinal targets in Appendix B.2, the discrete Wasserstein-1 formula and its normalized range in Appendix C.1, and our analysis of gradient paths under stop-gradient in Appendix E. We checked the references and related statements against the original papers. The code underwent smoke tests and was run to check its behavior. We take responsibility for the final manuscript, including its claims, references, code, and results.

ETHICS STATEMENT

We evaluate CoMoR on four benchmarks: CMU-MOSI, CMU-MOSEI, CH-SIMS, and CH-SIMS v2 (Zadeh et al., 2016; Bagher Zadeh et al., 2018; Yu et al., 2020; Liu et al., 2022). The recordings include people's faces and voices. We followed each dataset's terms of use and did not collect new recordings. Anyone reusing the data or sharing examples should follow the data providers' terms and protect the privacy of the people recorded.

REPRODUCIBILITY STATEMENT

The modality encoders and projectors are described in Appendix A, with diagnostic and routing details in Appendices B and D. Appendix F provides the exact prompts, sequence handling, and numerical decoding and parsing rules. Data sources, preprocessing, dataset splits, and evaluation metrics are documented in Appendix G.2. Training settings and random seeds are listed in Appendix G.1, and results across seeds are reported in Appendix H.1. Ablation controls and test-time token/frame missingness settings are given in Appendices G.4 and G.5. Derivations for the soft ordinal targets and normalized Wasserstein-1 distance, together with the gradient-flow analysis under stop-gradient, are provided in Appendices B.2, C, and E.

REFERENCES

AmirAli Bagher Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph. In Iryna Gurevych and Yusuke Miyao (eds.), *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2236–2246, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1208. URL https://aclanthology.org/P18-1208/.

Jinze Bai, Shuai Bai, Yunfei Chu, et al. Qwen technical report, 2023. URL https://arxiv.org/abs/2309.16609.

Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal machine learning: A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(2): 423–443, 2019. doi: 10.1109/TPAMI.2018.2798607. URL https://doi.org/10.1109/TPAMI.2018.2798607.

Grigor Bezirganyan, Sana Sellami, Laure Berti-Equille, and Sébastien Fournier. Multimodal learning with uncertainty quantification based on discounted belief fusion. In *Proceedings of The 28th International Conference on Artificial Intelligence and Statistics*, volume 258 of *Proceedings of Machine Learning Research*, pp. 3142–3150. PMLR, 2025. URL https://proceedings.mlr.press/v258/bezirganyan25a.html.

Yuchen Deng, Chang Sun, Hai-Tao Zheng, Feidiao Yang, and Yuxing Han. CAD: Conflict-aware decoding to mitigate cross-modal hallucinations in omnimodal large language models. arXiv preprint arXiv:2609.04247, 2026. URL https://arxiv.org/abs/2609.04247.

Benoit Dufumier, Javiera Castillo Navarro, Devis Tuia, and Jean-Philippe Thiran. What to align in multimodal contrastive learning? In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/108030643e640ac050e0ed5e6aace48f-Abstract-Conference.html.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, et al. The Llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

Zongbo Han, Changqing Zhang, Huazhu Fu, and Joey Tianyi Zhou. Trusted multi-view classification with dynamic evidential fusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(2):2551–2566, 2023. doi: 10.1109/TPAMI.2022.3171983. URL https://ieeexplore.ieee.org/document/9767662.

Devamanyu Hazarika, Roger Zimmermann, and Soujanya Poria. MISA: Modality-invariant and -specific representations for multimodal sentiment analysis. In *Proceedings of the 28th ACM International Conference on Multimedia*, MM '20, pp. 1122–1131, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450379885. doi: 10.1145/3394171.3413678. URL https://doi.org/10.1145/3394171.3413678.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9(8): 1735–1780, 1997. doi: 10.1162/neco.1997.9.8.1735.

Jiachen Hou, Zheyan Cheng, Chen Xing, Yushi Li, Nan Xiang, Shaoyu Cai, Jie Liu, and Yushan Pan. QDRF: Quantized and distilled robust fusion for incomplete data in multimodal sentiment analysis. *Information Sciences*, 758:123966, 2027. ISSN 0020-0255. doi: 10.1016/j.ins.2026.123966. URL https://www.sciencedirect.com/science/article/pii/S0020025526008972.

Zhongquan Jian, Wenhan Lv, Yanhao Chen, Guanran Luo, Wentao Qiu, Shaopan Wang, Bingbing Hu, and Qingqiang Wu. MDF: A modality-aware disentanglement and fusion framework for multimodal sentiment analysis. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40 (37):31292–31300, 2026. doi: 10.1609/aaai.v40i37.40392.

Xiaoran Kou, Jingyi Wu, Peng Sun, Yang Liu, and Hong Chen. Vision-guided text prompt tuning for multimodal sentiment analysis. arXiv preprint arXiv:2609.06497, 2026. URL https://arxiv.org/abs/2609.06497. Accepted to IEEE MMSP 2026.

Paul Liang, Chun Kai Ling, Yun Cheng, Alexander Obolenskiy, Yudong Liu, Rohan Pandey, Alex Wilf, Louis-Philippe Morency, and Russ Salakhutdinov. Multimodal learning without labeled multimodal data: Guarantees and applications. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (eds.), *International Conference on Learning Representations*, volume 2024, pp. 23072–23103, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/63f97950e32e6bdb4c6c9e31568b08c2-Paper-Conference.pdf.

J. Lin. Divergence measures based on the Shannon entropy. *IEEE Transactions on Information Theory*, 37(1):145–151, 1991. doi: 10.1109/18.61115.

Shanshan Lin, Yuesheng Wu, Chao Chen, Yizhe Yang, Zhihao Chen, Zexian Yang, and Xiangwen Liao. Multi-granularity sentiment integration for LLM-based multimodal sentiment analysis. arXiv preprint arXiv:2608.16201, 2026. URL https://arxiv.org/abs/2608.16201. Accepted to NLPCC 2026.

Yihe Liu, Ziqi Yuan, Huisheng Mao, Zhiyun Liang, Wanqiuyue Yang, Yuanzhe Qiu, Tie Cheng, Xiaoteng Li, Hua Xu, and Kai Gao. Make acoustic and visual cues matter: CH-SIMS v2.0 dataset and AV-Mixup consistent module. In *Proceedings of the 2022 International Conference on Multimodal Interaction (ICMI)*, ICMI '22, pp. 247–258. Association for Computing Machinery, 2022. doi: 10.1145/3536221.3556630. URL https://doi.org/10.1145/3536221.3556630.

Chunlei Meng, Jiabin Luo, Pengbin Feng, Zhenglin Yan, Chengyin Hu, Zhongxue Gan, and Chun Ouyang. Mitigating shared-private branch imbalance via dual-branch rebalancing for multimodal sentiment analysis, 2026. URL https://arxiv.org/abs/2604.25179.

Gabriel Peyré and Marco Cuturi. Computational optimal transport with applications to data sciences. *Foundations and Trends in Machine Learning*, 11(5-6):355–607, 02 2019. ISSN 1935-8237. doi: 10.1561/2200000073. URL https://doi.org/10.1561/2200000073.

Chengxuan Qian, Shuo Xing, Shawn Li, Yue Zhao, and Zhengzhong Tu. DecAlign: Hierarchical cross-modal alignment for decoupled multimodal representation learning. In *International Conference on Learning Representations (ICLR)*, 2026.

Wasifur Rahman, Md Kamrul Hasan, Sangwu Lee, AmirAli Bagher Zadeh, Chengfeng Mao, Louis-Philippe Morency, and Ehsan Hoque. Integrating multimodal information in large pretrained transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 2359–2369. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.acl-main.214. URL https://aclanthology.org/2020.acl-main.214/.

Team GLM, Aohan Zeng, Bin Xu, et al. ChatGLM: A family of large language models from GLM-130B to GLM-4 all tools. arXiv preprint arXiv:2406.12793, 2024. URL https://arxiv.org/abs/2406.12793.

THUIAR. MMSA: A unified framework for multimodal sentiment analysis. https://github.com/thuiar/MMSA, 2022. Results retrieved from results/result-stat.md.

Hugo Touvron, Louis Martin, Kevin Stone, et al. Llama 2: Open foundation and fine-tuned chat models, 2023. URL https://arxiv.org/abs/2307.09288.

Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J. Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. Multimodal transformer for unaligned multimodal language sequences. In Anna Korhonen, David Traum, and Lluís Màrquez (eds.), *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 6558–6569, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1656. URL https://aclanthology.org/P19-1656/.

Maria Tsimpoukelli, Jacob L Menick, Serkan Cabi, S. M. Ali Eslami, Oriol Vinyals, and Felix Hill. Multimodal few-shot learning with frozen language models. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 200–212. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/01b7575c38dac42f3cfb7d500438b875-Paper.pdf.

S. S. Vallender. Calculation of the Wasserstein distance between probability distributions on the line. *Theory of Probability & Its Applications*, 18(4):784–786, 1974. doi: 10.1137/1118101. URL https://doi.org/10.1137/1118101.

Pan Wang, Qiang Zhou, Yawen Wu, Tianlong Chen, and Jingtong Hu. DLF: Disentangled-language-focused multimodal sentiment analysis. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(20):21180–21188, 2025. doi: 10.1609/aaai.v39i20.35416.

Yan Wang, Fuyuan Cao, and Xingwang Zhao. Conflict-aware adaptive cross-reconstruction for multimodal sentiment analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15880–15889, June 2026. URL https://openaccess.thecvf.com/content/CVPR2026/papers/Wang_Conflict-Aware_Adaptive_Cross-Reconstruction_for_Multimodal_Sentiment_Analysis_CVPR_2026_paper.pdf.

Xiongye Xiao, Gengshuo Liu, Gaurav Gupta, Defu Cao, Shixuan Li, Yaxing Li, Tianqing Fang, Mingxi Cheng, and Paul Bogdan. Neuro-inspired information-theoretic hierarchical perception for multimodal learning. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/80790082a3b0e4fa9061730ee876f5ba-Abstract-Conference.html.

Yang Yang, Xunde Dong, and Yupeng Qiang. CLGSI: A multimodal sentiment analysis framework based on contrastive learning guided by sentiment intensity. In Kevin Duh, Helena Gomez, and Steven Bethard (eds.), *Findings of the Association for Computational Linguistics: NAACL 2024*, pp. 2099–2110, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-naacl.135. URL https://aclanthology.org/2024.findings-naacl.135/.

Yang Yang, Xunde Dong, and Yupeng Qiang. MSE-Adapter: A lightweight plugin endowing LLMs with the capability to perform multimodal sentiment analysis and emotion recognition. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(24):25642–25650, April 2025. doi: 10.1609/aaai.v39i24.34755. URL https://ojs.aaai.org/index.php/AAAI/article/view/34755.

Feng Yu, Xiangyu Wu, Yang Yang, and Jianfeng Lu. Multimodal classification via total correlation maximization. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/4ab56e2e7ea64ff654d28bc0fcaee108-Abstract-Conference.html.

Wenmeng Yu, Hua Xu, Fanyang Meng, Yilin Zhu, Yixiao Ma, Jiele Wu, Jiyun Zou, and Kaicheng Yang. CH-SIMS: A Chinese multimodal sentiment analysis dataset with fine-grained annotation of modality. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 3718–3727, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.343. URL https://aclanthology.org/2020.acl-main.343/.

Wenmeng Yu, Hua Xu, Ziqi Yuan, and Jiele Wu. Learning modality-specific representations with self-supervised multi-task learning for multimodal sentiment analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 10790–10797, 2021. doi: 10.1609/aaai.v35i12.17289.

Amir Zadeh, Rowan Zellers, Eli Pincus, and Louis-Philippe Morency. Multimodal sentiment intensity analysis in videos: Facial gestures and verbal messages. *IEEE Intelligent Systems*, 31(6):82–88, 2016. doi: 10.1109/MIS.2016.94. URL https://ieeexplore.ieee.org/document/7742221/.

Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Tensor fusion network for multimodal sentiment analysis. In Martha Palmer, Rebecca Hwa, and Sebastian Riedel (eds.), *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 1103–1114, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. doi: 10.18653/v1/D17-1115. URL https://aclanthology.org/D17-1115/.

Haoyu Zhang, Yu Wang, Guanghao Yin, Kejun Liu, Yuanyuan Liu, and Tianshu Yu. Learning language-guided adaptive hyper-modality representation for multimodal sentiment analysis. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 756–767, Singapore, December 2023a. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.49. URL https://aclanthology.org/2023.emnlp-main.49/.

Qingyang Zhang, Haitao Wu, Changqing Zhang, Qinghua Hu, Huazhu Fu, Joey Tianyi Zhou, and Xi Peng. Provable dynamic fusion for low-quality multimodal data. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 41753–41769. PMLR, 2023b. URL https://proceedings.mlr.press/v202/zhang23ar.html.

Xianbing Zhao, Lan Luo, Hengyang Lu, and Buzhou Tang. Prototype-as-Prompt: Multimodal sentiment prototypes endowing large language models the capability to perform multimodal sentiment analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 23010–23020, June 2026. URL https://openaccess.thecvf.com/content/CVPR2026/papers/Zhao_Prototype-as-Prompt_Multimodal_Sentiment_Prototypes_Endowing_Large_Language_Models_the_Capability_CVPR_2026_paper.pdf.

Aoqiang Zhu, Min Hu, Xiaohua Wang, Jiaoyun Yang, Yiming Tang, and Ning An. Multimodal invariant sentiment representation learning. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 14743–14755, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-256-5. doi: 10.18653/v1/2025.findings-acl.761. URL https://aclanthology.org/2025.findings-acl.761/.

Shihao Zou and Wei Wei. CoRiM: Conflict-driven risk minimization for dynamic multimodal fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 37821–37830, June 2026. URL https://openaccess.thecvf.com/content/CVPR2026/papers/Zou_CoRiM_Conflict-driven_Risk_Minimization_for_Dynamic_Multimodal_Fusion_CVPR_2026_paper.pdf.

A Detailed Architecture

This section details the architecture in Section 3. Unless otherwise stated, the configuration uses ChatGLM3-6B on CMU-MOSI and CH-SIMS, with $d = 4096$, $d_f = 256$, $N_p = 4$, and $K = 7$. Training settings are given in Appendix G.1.

A.1 Modality Encoders

Separate single-layer unidirectional LSTMs encode the pre-extracted audio and visual sequences. Their final hidden states are projected to 256-dimensional summaries without activation. The audio/visual hidden dimensions are $(64, 32)$ for MOSI and $(64, 64)$ for CH-SIMS; the output and the recurrent dropout are zero. Packed sequences use per-sample valid lengths clipped to $[1, L]$ for padded width L, not configured maximum lengths.

The text is truncated or padded to 50 tokens, excluding any automatically added special tokens. Only token IDs and attention masks are used; the stored segment-ID channel is unused. The diagnostic branch applies a $4096 \to 1$ attention-score map and $4096 \to 256$ projection with GELU (Section 3.1). For valid indices $\mathcal{T}$ and frozen vocabulary embeddings $E_i^t \in \mathbb{R}^d$, pooling is

$$\begin{aligned} \alpha_i &= \frac{\exp(w_{\text{pool}}^\top E_i^t)}{\sum_{j\in\mathcal{T}} \exp(w_{\text{pool}}^\top E_j^t)}, \qquad i \in \mathcal{T}, \\ f^t &= \text{GELU}\left(W_t \sum_{i\in\mathcal{T}} \alpha_i E_i^t + c_t \right). \end{aligned} \tag{15}$$

Here w_{pool}, W_t, and c_t are trainable. Padding positions receive zero pooling weight. The attention-score layer's shared scalar bias cancels under softmax and is therefore omitted from Eq. (15). At least one valid text token is required before using modality-availability masking. This pooling operation is applied prior to contextual encoding. For fixed word embeddings and their corresponding valid masks, the pooling output is invariant to token order. This limitation impacts the diagnostic text branch: joint prediction preserves the original ordered sequence of text embeddings.

A.2 Independent Multi-scale Projectors

Each modality has an independent instance of the multi-scale projector adapted from MSE-Adapter (Yang et al., 2025). For a summary $f \in \mathbb{R}^{256}$, branch $q \in \{1, 2, 3\}$ computes

$$v_q = W_{q,2}\,\text{GELU}(W_{q,1} f + \beta_{q,1}) + \beta_{q,2} \in \mathbb{R}^{256}. \tag{16}$$

The hidden widths, in branch order, are $(512, 128, 256)$ for $d = 4096$. There is no dropout or final activation inside these branches. Stacking the branch outputs produces a 256×3 array. A convolution with one input channel, one output channel, and a 1×3 kernel integrates the scale axis. Equivalently,

$$v = \sum_{q=1}^{3} \kappa_q v_q + \kappa_0 \mathbf{1}_{256}, \qquad h = W_h v + \beta_h \in \mathbb{R}^d. \tag{17}$$

Here, $W_h \in \mathbb{R}^{d\times 256}$. A shared-coordinate linear map from one channel to $N_p = 4$ channels produces the token slots:

$$A_m(f)_{j,:} = \eta_j h + \beta_j \mathbf{1}_d, \qquad j = 1, \ldots, N_p. \tag{18}$$

Thus, these four slots are essentially different affine transformations applied to the same summary vector, rather than separate encodings of independent time intervals; the varying scales represent differences in MLP hidden layer width, not changes in temporal resolution.

A learned modality embedding $e^m \in \mathbb{R}^d$ is added to every slot, as described in Section 3.1:

$$H^m = A_m(f^m) + \mathbf{1}_{N_p}(e^m)^\top \in \mathbb{R}^{N_p\times d}. \tag{19}$$

The three modality embeddings are initialized from $\mathcal{N}(0, 0.02^2)$. Projector and modality-embedding parameters are not shared across modalities. Text pseudo-tokens are used only for diagnosis. For complete inputs, the joint prefix includes four audio and four visual pseudo-tokens, together with the original text and task prompt.

B Diagnostic Prediction and Ordered Targets

B.1 Shared Diagnostic Interface

Each modality uses the diagnostic input format described in Section 3.2. The markers `<Multimodal>` and `</Multimodal>` are ordinary tokenized strings, while BOS comes from the tokenizer. The shared head takes the hidden state at the last valid prompt position.

Let ψ denote the head parameters in Eq. (2). For ChatGLM3-6B, the head applies 4096-dimensional layer normalization, a $4096 \to 256$ linear map, GELU, dropout, and a $256 \to 7$ linear map, with logits converted to float32 before softmax. Within each dataset, all modalities share this head, frozen backbone, prompt, and anchor support, without temperature fitting or calibration.

All branches use the sample-level target y rather than modality-specific labels; thus, their predictions cannot identify the correct modality or its independently annotated sentiment. For a microbatch of B samples, the diagnostic backbone processes B text sequences and $2B$ audio/visual sequences independently, without cross-modal attention or shared hidden activations.

B.2 Soft Ordinal Supervision

The soft targets follow Eq. (11). The CMU-MOSI anchors are $(-3, -2, -1, 0, 1, 2, 3)$, and the CH-SIMS anchors are $(-1, -2/3, -1/3, 0, 1/3, 2/3, 1)$. For the clipped target $\bar{y} \in [a_k, a_{k+1}]$, their expectation is

$$\sum_{j=1}^{K} a_j q_j(y) = \frac{a_k(a_{k+1} - \bar{y}) + a_{k+1}(\bar{y} - a_k)}{a_{k+1} - a_k} = \bar{y}. \tag{20}$$

The target therefore preserves the clipped continuous value in expectation. Diagnostic supervision uses this continuous value, whereas the generated training response is rounded to one decimal place.

C Properties of the Diagnostic Statistics

C.1 Normalized Wasserstein-1

On a common ordered support, each cumulative distribution is constant on $[a_k, a_{k+1})$. The one-dimensional identity $W_1(p, q) = \int_{\mathbb{R}} |F_p(x) - F_q(x)|\, dx$ therefore yields

$$W_1(p, q) = \sum_{k=1}^{K-1} (a_{k+1} - a_k) \left| \sum_{j=1}^{k} (p_j - q_j) \right|. \tag{21}$$

This standard finite-support identity (Peyré & Cuturi, 2019; Vallender, 1974) requires no transport approximation. Cumulative differences lie in $[-1, 1]$ and anchor gaps sum to $a_K - a_1$, giving $0 \le W_1(p, q)/(a_K - a_1) \le 1$, with equality at one for opposite endpoint masses.

The implementation renormalizes nonnegative finite probability masses and sums the absolute cumulative probability differences over the first $K-1$ anchors, weighted by the normalized anchor gaps. Computations in float16/bfloat16 are promoted to float32, and the result is clipped to $[0,1]$ to account for numerical rounding errors. Sinkhorn iterations are not used.

C.2 Predictive Entropy and Redundancy

Dividing categorical entropy by its maximum $\log K$, attained at uniformity, gives Eq. (5). The ideal convention is $0 \log 0 = 0$. In each of the two factors of the addend, achieve a probability of 10^{-8} without requiring renormalization.

Wasserstein distance and entropy are deterministic summaries of the diagnostic distributions; they add no information beyond those distributions and do not establish modality correctness or reliability.

D Routing and Availability Masking

D.1 Descriptor Construction and Availability Masking

We use the descriptor S defined in Eq. (6), with dimension $3K+9=30$. Table 1 uses $(b^t, b^a, b^v) = (1,1,1)$; otherwise, $b^m \in \{0,1\}$ and $n_b = \sum_{m\in\mathcal{M}} b^m \geq 1$.

Absent probability blocks become $\mathbf{1}_K/K$. Entropy and pairwise disagreement are multiplied by b^m and $b^m b^n$, respectively, so missing inputs contribute zero summaries. Availability bits distinguish this missing-value sentinel from confidence.

The $30 \to 64 \to 3$ router uses GELU and dropout between linear layers and receives detached $\mathrm{sg}(S)$ (Eq. (7)). We initialize the final-layer weights and biases to zero. Modality ablations normalize weights over retained modalities, excluded ones will be assigned zero weight. We then replace the scaling factor 3 by n_b, the number of retained modalities. Auxiliary loss averages available sample–modality pairs per mini-batch; all-absent samples are rejected.

The joint mask excludes absent audio/visual blocks and multiplies the raw-text mask by b^t. Masked gaps are removed in valid-token order before left padding. Text pooling requires nonempty pre-mask text. Absent audio/visual diagnoses still process zeroed pseudo-tokens before statistics are replaced; masking neither removes all diagnostic computation nor establishes missing-modality robustness.

D.2 Uniform-Routing Scale and Text Asymmetry

At initialization, present modalities have equal logits, so $W^m = 1/n_b$ for $b^m = 1$. The audio and visual coefficients therefore satisfy

$$n_b W^m = 1, \qquad m \in \{a, v\} \text{ with } b^m = 1. \tag{22}$$

Uniform routing therefore leaves each present audio/visual pseudo-token block, including its modality embedding, unchanged. The original text remains unscaled.

E Gradient Flow

Table 4: Gradient paths from each loss to the trainable modules.

Module	$\mathcal{L}_{\text{aux}}$	$\mathcal{L}_{\text{task}}$
Text diagnostic pool/projection/projector	Yes	No
Audio encoder/projector	Yes	Yes
Visual encoder/projector	Yes	Yes
Shared diagnostic head	Yes	No
Router	No	Yes

The frozen LLM transmits gradients to continuous inputs without updating its parameters. Stop-gradient blocks task-loss gradients through the diagnostic descriptors, while the router learns from the task loss. Audio/visual encoders and projectors receive auxiliary-loss gradients through diagnosis and task-loss gradients through the scaled joint inputs. Modality embeddings follow the gradient paths of their corresponding projectors. Shared parameter updates can still alter subsequent diagnoses.

F Prompt Templates and Numerical Decoding

F.1 Exact Templates

The exact MOSI and CH-SIMS prompts are shown below and are tokenized without automatically added special tokens. Line wrapping is typographical. Diagnostic prompts are shared across modalities within each dataset. MOSEI reuses both MOSI prompts, and CH-SIMS v2 reuses both CH-SIMS prompts, for both diagnostic and final task prediction.

MOSI diagnostic prompt.

> Predict this modality's sentiment intensity from -3 to +3.

MOSI final task prompt.

> Please predict the sentiment intensity of the above multimodal content in the range [-3.0, 3.0]. response: The sentiment is

CH-SIMS diagnostic prompt.

> 请预测该模态的情感强度，范围为-1到+1。

English translation: Predict this modality's sentiment intensity from -1 to +1.

CH-SIMS final task prompt.

> 请对上述多模态内容的情感强度进行预测，范围在[-1.0, 1.0]之间。响应: 情感为

English translation: Please predict the sentiment intensity of the above multimodal content in the range [-1.0, 1.0]. Response: The sentiment is

The English translations are provided only for explanation; the Chinese strings are used as model inputs.

The joint input order follows Eq. (10): BOS, opening marker, audio pseudo-tokens, visual pseudo-tokens, closing marker, ordered raw text, and the task prompt. Notably, the diagnostic prompt is not included in this joint prefix, whose marker and prompt lengths depend on the tokenizer assets that are used.

F.2 Sequence Handling and Output Rules

Training targets are clipped to the dataset range and formatted with an explicit sign and one decimal place without appended EOS supervision. Causal loss is averaged over response tokens per microbatch, with prefix and padding positions ignored using index -100.

Greedy decoding generates at most four new tokens, stopping when all sequences emit EOS; finished sequences are padded with EOS. Key-value caching is disabled.

For evaluation, U+2013 and U+2212 are converted to ASCII minus, and the first integer or decimal substring with an optional sign is parsed. Parsing failures and nonfinite or out-of-range values are assigned score zero; out-of-range values are not clipped.

G Experimental Configuration and Computational Accounting

G.1 Main Experimental Setup and Training Protocol

The main evaluation covers CMU-MOSI, CMU-MOSEI, CH-SIMS, and CH-SIMS v2. ChatGLM3-6B is the frozen backbone unless otherwise stated. The main results for CoMoR are averaged over three random seeds: 1111, 1113, and 1115. MSE-Adapter results are averaged over five random seeds: 1111, 2222, 3333, 4444, and 5555. Standard deviations are computed as sample standard deviations. Checkpoints were selected using MAE.

CoMoR is trained with AdamW for 40 epochs. The microbatch size is 4, with gradient accumulation over 4 steps, giving an effective batch size of 16. We use an AdamW epsilon of 10^{-4} and a gradient clipping threshold of 1.0. The learning rate is warmed up over the first 10% of optimizer updates and then decayed using a cosine schedule.

The objective combines the task loss $\mathcal{L}_{\text{task}}$ with the auxiliary diagnostic loss, using $\lambda = 0.3$. Weight decay is 0.01 for all four datasets. The LLM parameters remain frozen throughout training. Table 5 gives the dataset-specific settings.

Table 5: Dataset-specific training settings. LR denotes learning rate. The diagnostic head and router share the listed learning rate and dropout.

Dataset	Adapter LR	Head/router LR	Dropout
CMU-MOSI	3×10^{-4}	3×10^{-4}	0.1
CMU-MOSEI	5×10^{-3}	1×10^{-3}	0.1
CH-SIMS	1×10^{-3}	1×10^{-3}	0.0
CH-SIMS v2	1×10^{-4}	1×10^{-4}	0.0

The adapter parameter group includes the trainable text, audio, and visual adaptation modules and the modality embeddings. The diagnostic head and router form a separate parameter group. All trainable modules are optimized jointly.

G.2 Dataset and Preprocessing Details

We use CMU-MOSI (Zadeh et al., 2016), CMU-MOSEI (Bagher Zadeh et al., 2018), CH-SIMS (Yu et al., 2020), and CH-SIMS v2 (Liu et al., 2022), retaining the train/validation/test assignments in the processed files. Table 6 lists split sizes, target ranges, and input configurations for ChatGLM3-6B. Feature dimensions, cached sequence lengths, and LSTM hidden dimensions are given in audio–vision order; each sample retains its own valid sequence lengths. We use only the supervised partition of CH-SIMS v2. All diagnostic branches use the sample-level multimodal label, including on the CH-SIMS datasets, which also provide modality-specific labels.

Table 6: Input configurations for the four datasets. Split sizes refer to the processed data files used in our experiments.

Setting	MOSI	MOSEI	CH-SIMS	CH-SIMS v2
Target range	$[-3, 3]$	$[-3, 3]$	$[-1, 1]$	$[-1, 1]$
AV feature dimensions	$(5, 20)$	$(74, 35)$	$(33, 709)$	$(25, 177)$
Cached AV lengths	$(375, 500)$	$(500, 375)$	$(400, 55)$	$(925, 232)$
LSTM hidden dimensions	$(64, 32)$	$(64, 32)$	$(64, 64)$	$(64, 64)$
Training samples	1284	16326	1368	2722
Validation samples	229	1871	456	647
Test samples	686	4659	457	1034

All four datasets use seven equally spaced sentiment anchors over their target ranges, four diagnostic pseudo-tokens per modality, and single-layer audio/visual LSTMs with zero LSTM dropout and outputs projected to 256 dimensions.

We tokenize the supplied raw transcripts with the backbone tokenizer, using truncation and padding to 50 tokens without automatically added special tokens. This limit applies to the transcript alone, before adding prompts, markers, and pseudo-tokens. Audio and visual features remain unaligned. Their recorded valid lengths are used for packed LSTM encoding, with lengths bounded between 1 and the stored sequence length.

The loader replaces negative infinities in audio features and NaNs in visual features with zero. For CH-SIMS, values with absolute magnitude above 10000 in visual dimensions 288, 289, and 290 (zero-based indices) are also replaced by zero. This additional cleaning step is not applied to CH-SIMS v2. No additional feature normalization is applied.

Data sources. We use the supplied transcripts, sentiment labels, and pre-extracted audio and visual features. Features for CMU-MOSI, CMU-MOSEI, and CH-SIMS were downloaded through the link in the MMSA repository.[1] A public copy of the CH-SIMS v2 videos and transcripts is available in THUIAR's MMLA repository.[2]

Evaluation metrics. We report mean absolute error (MAE) and Pearson correlation (Corr) for regression, and binary accuracy (Acc-2) and class-support-weighted F1 for classification. Acc-2 and F1 are percentages, with scores greater than zero classified as positive and all others as nonpositive. Classification excludes zero-label samples on CMU-MOSI and CMU-MOSEI. CH-SIMS and CH-SIMS v2 retain zero-label samples and clip predictions and targets to $[-1, 1]$ for classification. MAE and Corr use all evaluation samples.

G.3 Computational Accounting

Table 7 lists the trainable parameter counts, excluding the frozen backbone. For CMU-MOSI, the total of 6,715,695 parameters corresponds to the 6.72 million reported in Figure 4.

Table 7: Analytical counts of trainable parameters for the specified $d = 4096$, $N_p = 4$ architecture.

Module	MOSI	CH-SIMS
Three modality projectors	4,539,300	4,539,300
Text pooling and projection	1,052,929	1,052,929
Three modality embeddings	12,288	12,288
Shared diagnostic head	1,058,823	1,058,823
Router	2,179	2,179
Audio encoder and projection	34,816	41,984
Visual encoder and projection	15,360	215,040
Total trainable	6,715,695	6,922,543

Each training microbatch uses two diagnostic backbone calls (one for text and one for audio/vision), followed by one joint prediction call. At inference, diagnostics are computed once per sample, and each decoding step reevaluates the growing joint prefix with the key-value cache disabled.

G.4 Ablation and Additional Analysis Settings

Unless otherwise stated, all modules, losses, and training settings follow Appendix G.1.

Alternative backbones. We use `THUDM/chatglm3-6b-base`, `Qwen/Qwen-1_8B`, `meta-llama/Llama-3.2-3B`, and `meta-llama/Llama-2-7b-hf`, with hidden dimensions 4096, 2048, 3072, and 4096, respectively. Interfaces are resized to each backbone's hidden dimension; the diagnostic head retains 256 hidden units and seven output classes, and the router retains its $30 \to 64 \to 3$ architecture. Within each dataset, prompt text and input ordering

[1] `https://github.com/thuiar/MMSA`

[2] `https://huggingface.co/datasets/THUIAR/MMLA-Datasets/tree/main/CH-SIMSv2.0`

are shared. Each backbone uses its own tokenizer and special tokens, with EOS used for padding when no separate padding token exists. Greedy decoding generates at most four new tokens, using the parsing rules in Appendix F.2.

Ablation controls. Component and routing ablations use the CMU-MOSI and CH-SIMS settings in Table 5. All variants retain the frozen LLM for final prediction and the auxiliary diagnostic loss with $\lambda = 0.3$.

Removing the diagnostic LLM. For each modality, we mean-pool its four pseudo-token embeddings and feed the resulting vector directly to the shared diagnostic head, bypassing the diagnostic LLM.

Removing routing inputs. We remove conflict inputs, entropy inputs, or both by setting their coordinates to zero while keeping the router input dimension at 30. For w/o D, we remove diagnostic distributions from the router input but retain their conflict and entropy summaries and the availability indicators.

Uniform and learned global gates. Uniform gates bypass the routing MLP and assign weights of $(1/3, 1/3, 1/3)$ for complete inputs, giving unit scaling to audio and visual pseudo-tokens. Learned global gates replace the routing MLP with three trainable logits shared across samples, initialized to zero and normalized by softmax. They use the head learning rate and receive gradients from the task loss.

Independent diagnostic heads. We replace the shared diagnostic head with three separate heads of the same architecture. Their initial parameters are copied from the shared head, after which they are trained independently.

Alternative conflict measures. Each alternative replaces only the three conflict coordinates. Hidden cosine conflict uses $(1 - \cos(h^m, h^n))/2$, where h^m and h^n are the diagnostic hidden vectors from the same forward pass. Normalized Jensen–Shannon (JS) divergence (Lin, 1991) uses $\mathrm{JS}(P^m, P^n)/\log 2$, without a square root.

Feature-driven routing. Each 256-dimensional modality summary is detached before a separate trainable linear layer maps it to ten dimensions. The concatenated outputs replace the diagnostic descriptor at the input of the $30 \to 64 \to 3$ router, adding $3(256 \times 10 + 10) = 7{,}710$ parameters. Task-loss gradients reach the added projections and router, but do not propagate through this routing path into the upstream feature extraction modules.

G.5 Test-Time Token/Frame Missingness

We assess random token/frame missingness on CMU-MOSI and CH-SIMS, a form of incomplete-input evaluation also considered by QDRF (Hou et al., 2027), using the three CoMoR checkpoints described in Appendix G.1. The target rate $r \in \{0, 0.1, \ldots, 0.9\}$ applies to text, audio, and vision. Each eligible position is selected with probability r; at $r = 0$, inputs remain unchanged.

Eligible text positions exclude padding and protected tokenizer special tokens; UNK and PAD IDs are not protected as special tokens. Selected text tokens become UNK (ID 0), while selected audio/visual frames have their feature vectors zeroed. Attention masks and valid sequence lengths remain unchanged. Prompts, BOS, and multimodal wrapper markers are added after corruption and remain intact. All eligible positions may be corrupted, and the router always receives `presence=[1,1,1]`.

Masks use NumPy `RandomState(1111)`, reset separately to the same seed for each modality, so modality masks are not independent. Masks are shared across checkpoints, and random numbers are reused across rates, making lower-rate masks subsets of higher-rate masks. The same checkpoints are evaluated at every rate, with one test-set pass per checkpoint and rate and no weight updates. Means and sample standard deviations describe variation across training seeds under fixed masks. Metrics follow Appendix G.2, and results appear in Appendix H.3.

H Further Experimental Results

H.1 Results Across Random Seeds

Table 8 reports mean ± sample standard deviation across the model-specific training seeds listed in Appendix G.1, using the metrics defined in Appendix G.2.

Table 8: Results across random seeds, reported as mean ± sample standard deviation. Acc-2 and F1 are percentages; the CH-SIMS datasets retain zero-valued targets.

CMU-MOSI				
Model	MAE ↓	Corr ↑	Acc-2 ↑	F1 ↑
MSE-Adapter	0.672 ± 0.012	0.834 ± 0.010	88.38 ± 1.68	88.30 ± 1.78
CoMoR	0.581 ± 0.013	0.864 ± 0.003	89.28 ± 0.32	89.21 ± 0.34
CMU-MOSEI				
Model	MAE ↓	Corr ↑	Acc-2 ↑	F1 ↑
MSE-Adapter	0.516 ± 0.018	0.786 ± 0.020	86.61 ± 0.63	86.31 ± 0.91
CoMoR	0.493 ± 0.003	0.796 ± 0.003	86.88 ± 0.56	86.72 ± 0.43
CH-SIMS				
Model	MAE ↓	Corr ↑	Acc-2 ↑	F1 ↑
MSE-Adapter	0.395 ± 0.022	0.626 ± 0.027	79.47 ± 2.37	77.39 ± 3.48
CoMoR	0.349 ± 0.009	0.684 ± 0.012	81.98 ± 1.76	80.82 ± 2.30
CH-SIMS v2				
Model	MAE ↓	Corr ↑	Acc-2 ↑	F1 ↑
MSE-Adapter	0.306 ± 0.023	0.705 ± 0.032	81.43 ± 2.15	81.30 ± 2.19
CoMoR	0.271 ± 0.010	0.752 ± 0.017	83.88 ± 0.49	83.76 ± 0.47

H.2 Ablation Results Across Random Seeds

Tables 9, 10, and 11 report mean ± sample standard deviation across the three CoMoR training seeds listed in Appendix G.1. Metrics follow Appendix G.2.

H.3 Performance under Random Token/Frame Missingness

Table 12 reports the evaluation defined in Appendix G.5. Comparisons use each dataset's 0% retest; masks are shared across checkpoints and nested across rates, with all modality-availability indicators enabled.

Performance generally deteriorates as missingness increases on both datasets. At 90% missingness, CH-SIMS retains 69.00% Acc-2 while Corr falls to 0.038, showing that binary accuracy alone does not capture the extent of degradation. CMU-MOSI also exhibits greater F1 variation across training seeds under severe missingness. These results characterize sensitivity under the specified fixed-mask protocol.

Table 9: Component ablations. DL: diagnostic LLM; D denotes diagnostic distribution inputs; C: explicit conflict inputs; U: entropy inputs. Values are mean ± sample standard deviation over three seeds. Best means per dataset are bold. Acc-2 and F1 are percentages.

CMU-MOSI				
Variant	MAE ↓	Corr ↑	Acc-2 ↑	F1 ↑
CoMoR (full)	$\mathbf{0.581} \pm 0.013$	$\mathbf{0.864} \pm 0.003$	$\mathbf{89.28} \pm 0.32$	$\mathbf{89.21} \pm 0.34$
w/o DL	0.599 ± 0.010	0.858 ± 0.005	87.70 ± 2.18	87.46 ± 2.37
w/o D	0.590 ± 0.006	0.853 ± 0.006	88.48 ± 1.00	88.37 ± 1.03
w/o C	0.589 ± 0.003	0.860 ± 0.004	88.41 ± 1.33	88.27 ± 1.39
w/o U	0.589 ± 0.008	0.859 ± 0.000	88.16 ± 0.98	88.05 ± 1.11
w/o C and U	0.600 ± 0.008	0.853 ± 0.008	87.70 ± 1.07	87.60 ± 1.20
CH-SIMS				
Variant	MAE ↓	Corr ↑	Acc-2 ↑	F1 ↑
CoMoR (full)	$\mathbf{0.349} \pm 0.009$	$\mathbf{0.684} \pm 0.012$	$\mathbf{81.98} \pm 1.76$	$\mathbf{80.82} \pm 2.30$
w/o DL	0.358 ± 0.011	0.658 ± 0.019	80.53 ± 0.58	78.89 ± 1.12
w/o D	0.357 ± 0.018	0.671 ± 0.026	80.67 ± 2.97	78.66 ± 1.68
w/o C	0.357 ± 0.005	0.668 ± 0.009	80.53 ± 0.76	78.73 ± 1.71
w/o U	0.369 ± 0.006	0.644 ± 0.007	80.38 ± 0.33	79.15 ± 0.65
w/o C and U	0.356 ± 0.011	0.667 ± 0.018	80.45 ± 1.28	78.75 ± 2.75

Table 10: Modality ablations. T: text; A: audio; V: vision. Values are mean ± sample standard deviation over three seeds. Best means per dataset are bold. Acc-2 and F1 are percentages.

CMU-MOSI				
Variant	MAE ↓	Corr ↑	Acc-2 ↑	F1 ↑
T + A	0.601 ± 0.039	0.858 ± 0.009	$\mathbf{89.38} \pm 1.08$	$\mathbf{89.33} \pm 1.09$
T + V	0.596 ± 0.016	0.856 ± 0.005	88.06 ± 0.76	87.91 ± 0.79
A + V	1.604 ± 0.128	0.148 ± 0.035	56.96 ± 1.23	57.06 ± 1.36
T + A + V	$\mathbf{0.581} \pm 0.013$	$\mathbf{0.864} \pm 0.003$	89.28 ± 0.32	89.21 ± 0.34
CH-SIMS				
Variant	MAE ↓	Corr ↑	Acc-2 ↑	F1 ↑
T + A	0.360 ± 0.017	0.657 ± 0.019	80.82 ± 1.28	79.53 ± 1.53
T + V	0.361 ± 0.002	0.667 ± 0.009	$\mathbf{82.35} \pm 0.33$	$\mathbf{81.28} \pm 0.22$
A + V	0.564 ± 0.018	0.124 ± 0.046	69.37 ± 0.00	56.84 ± 0.00
T + A + V	$\mathbf{0.349} \pm 0.009$	$\mathbf{0.684} \pm 0.012$	81.98 ± 1.76	80.82 ± 2.30

Table 11: Diagnostic and routing design choices. Values are mean ± sample standard deviation over three seeds. Best means per dataset are bold. F1 is a percentage.

CMU-MOSI		
Variant	MAE ↓	F1 ↑
CoMoR (full)	$\mathbf{0.581} \pm 0.013$	89.21 ± 0.34
Independent diagnostic heads	0.605 ± 0.018	86.65 ± 0.98
Hidden cosine conflict	0.595 ± 0.005	88.16 ± 1.03
JS divergence	0.591 ± 0.009	$\mathbf{89.65} \pm 1.06$
Uniform gates	0.600 ± 0.004	88.52 ± 1.07
Learned global gates	0.600 ± 0.009	87.25 ± 2.10
Feature-driven routing	0.597 ± 0.005	87.35 ± 1.30
CH-SIMS		
Variant	MAE ↓	F1 ↑
CoMoR (full)	$\mathbf{0.349} \pm 0.009$	80.82 ± 2.30
Independent diagnostic heads	0.364 ± 0.011	78.18 ± 2.93
Hidden cosine conflict	0.360 ± 0.007	78.33 ± 1.42
JS divergence	0.358 ± 0.003	80.51 ± 1.77
Uniform gates	0.363 ± 0.009	76.46 ± 1.17
Learned global gates	0.377 ± 0.008	79.33 ± 2.35
Feature-driven routing	0.354 ± 0.002	$\mathbf{81.39} \pm 0.43$

Table 12: Test-time token/frame missingness on CMU-MOSI and CH-SIMS. Mean ± sample standard deviation across training seeds 1111, 1113, and 1115, using shared masks generated with seed 1111. Acc-2 and F1 are percentages. Average includes all ten rates (0–90%).

CMU-MOSI				
Target rate	Acc-2 ↑	F1 ↑	Corr ↑	MAE ↓
0%	89.28 ± 0.32	89.21 ± 0.34	0.864 ± 0.003	0.581 ± 0.013
10%	83.18 ± 1.99	82.95 ± 2.13	0.759 ± 0.013	0.745 ± 0.010
20%	77.19 ± 1.93	76.60 ± 2.29	0.651 ± 0.006	0.905 ± 0.011
30%	72.46 ± 0.69	71.29 ± 1.46	0.552 ± 0.016	1.041 ± 0.014
40%	71.19 ± 0.76	69.30 ± 1.81	0.490 ± 0.006	1.126 ± 0.008
50%	65.50 ± 1.02	61.96 ± 0.82	0.362 ± 0.017	1.242 ± 0.003
60%	60.01 ± 1.67	54.74 ± 1.65	0.235 ± 0.009	1.329 ± 0.004
70%	59.25 ± 1.04	52.45 ± 2.98	0.161 ± 0.008	1.364 ± 0.003
80%	58.43 ± 0.89	50.34 ± 4.85	0.100 ± 0.018	1.387 ± 0.004
90%	58.94 ± 1.02	50.03 ± 5.20	0.101 ± 0.019	1.387 ± 0.004
Average	69.54 ± 0.16	65.89 ± 2.20	0.427 ± 0.004	1.111 ± 0.003
CH-SIMS				
Target rate	Acc-2 ↑	F1 ↑	Corr ↑	MAE ↓
0%	81.98 ± 1.76	80.82 ± 2.30	0.684 ± 0.012	0.349 ± 0.009
10%	78.56 ± 1.53	75.84 ± 2.88	0.634 ± 0.033	0.387 ± 0.019
20%	75.71 ± 1.43	71.86 ± 2.94	0.560 ± 0.040	0.434 ± 0.019
30%	74.33 ± 0.46	69.38 ± 1.41	0.513 ± 0.029	0.461 ± 0.014
40%	72.50 ± 0.55	66.18 ± 0.91	0.424 ± 0.005	0.500 ± 0.008
50%	71.12 ± 1.33	63.47 ± 1.89	0.355 ± 0.016	0.533 ± 0.006
60%	70.90 ± 0.95	62.42 ± 1.10	0.304 ± 0.017	0.550 ± 0.001
70%	69.51 ± 1.32	59.99 ± 1.09	0.210 ± 0.044	0.581 ± 0.007
80%	69.07 ± 1.08	58.19 ± 0.92	0.123 ± 0.070	0.595 ± 0.009
90%	69.00 ± 0.33	57.16 ± 0.73	0.038 ± 0.044	0.602 ± 0.002
Average	73.27 ± 1.07	66.53 ± 1.62	0.385 ± 0.031	0.499 ± 0.009

H.4 PREDICTED-CONFLICT LEVELS

We analyze the CMU-MOSI run with seed 1115. For each sample, the full model provides the three pairwise conflict values, whose arithmetic mean is $c = (C^{ta} + C^{tv} + C^{av})/3$. We set q_1 and q_2 to the $1/3$ and $2/3$ quantiles of these scores on the validation set, using linear interpolation. With these thresholds fixed, test samples are assigned to low ($c \leq q_1$), medium ($q_1 < c \leq q_2$), or high ($c > q_2$) conflict groups. Both models in Table 13 use the groups determined by the full model; test-set group sizes need not be equal. Relative to the variant without conflict inputs, CoMoR yields the largest MAE reduction in the high-conflict group, a smaller reduction in the medium group, and a slight increase in the low-conflict group.

Table 13: MAE by predicted-conflict stratum for the CMU-MOSI run with seed 1115. N denotes the number of samples. Δ is the MAE without conflict inputs minus the full-model MAE, computed from unrounded values.

Conflict	N	w/o conflict	CoMoR	$\Delta \uparrow$
Low	267	0.594	0.606	-0.012
Medium	204	0.595	0.576	$+0.019$
High	215	0.571	0.532	$+0.039$
All	686	0.587	0.574	$+0.013$