CONTRIFUSE: CONTRIBUTION-AWARE FUSION OF AUXILIARY DESCRIPTIONS FOR MULTIMODAL SENTIMENT ANALYSIS

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal Sentiment Analysis (MSA) integrates textual, acoustic, and visual signals to infer sentiment polarity and intensity. Existing methods typically encode inputs from different modalities into feature representations and perform sentiment prediction through cross-modal fusion. Recent studies further introduce auxiliary information, such as descriptions of audio-visual content and contextual knowledge, to enrich the semantic information contained in modality representations. Existing approaches mainly use such auxiliary textual information to augment or align modality representations, followed by joint, gated, or routed fusion for sentiment prediction. However, auxiliary descriptions from different modalities do not contribute equally to sentiment prediction; their usefulness depends not only on the corresponding original modality information, but also on auxiliary descriptions from other modalities. To address this issue, we propose CONTRIFUSE, a multimodal sentiment analysis framework that adaptively regulates the contribution of each modality's auxiliary descriptions before integrating them with the original modality features. First, CONTRIFUSE constructs an auxiliary description representation for each modality from descriptions generated by multiple large language models. The resulting auxiliary representations are then mapped into shared and modality-specific spaces, allowing subsequent regulation to consider both information shared across modalities and information specific to each modality. CONTRIFUSE further combines the original modality features, auxiliary description representations, and semantic similarities among descriptions from different modalities to determine the extent to which each modality's auxiliary description should contribute to the current prediction. Finally, CONTRIFUSE fuses the modulated auxiliary description representations with the corresponding modality features to predict sentiment polarity and intensity. Experiments on four widely used English and Chinese sentiment analysis benchmarks evaluate CONTRIFUSE using both regression and classification metrics and demonstrate its effectiveness.

1 INTRODUCTION

Multimodal Sentiment Analysis (MSA) infers sentiment polarity and intensity from linguistic, acoustic, and visual signals (Zadeh et al., 2016; 2018). Multimodal affect modeling supports human–computer interaction (Jaimes & Sebe, 2007), service-call sentiment analysis (Jia, 2020), and emotion recognition in video learning (Zhang et al., 2025). Existing MSA methods learn and fuse original-modality representations through high-order interaction, cross-modal temporal modeling, or shared–private learning (Zadeh et al., 2017; Liu et al., 2018; Tsai et al., 2019; Hazarika et al., 2020; Han et al., 2021; Yu et al., 2021; Zhang et al., 2023). While linguistic sentiment is often explicit, non-verbal affective cues remain implicit in acoustic and visual features.

To make such information more explicit, recent work introduces auxiliary information alongside the original multimodal features. TextMI and DEVA convert acoustic and visual cues into natural-language descriptions (Hasan et al., 2023; Wu et al., 2025), while WisdoM introduces contextual world knowledge to complement the observed multimodal content (Wang et al., 2024). Other methods study how auxiliary information can be incorporated into prediction: MSE-Adapter connects multimodal inputs with LLM reasoning (Yang et al., 2025b); TEXT aligns generated explanations

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.

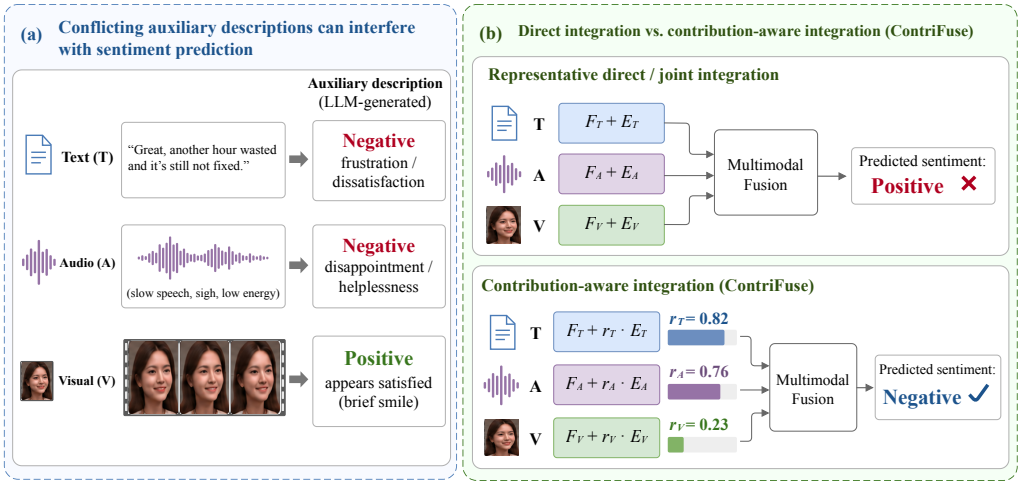


Figure 1: Contribution-aware fusion. Auxiliary descriptions can vary in usefulness across samples and modalities. CONTRIFUSE regulates their contributions before source-feature integration. F_m: source-modality feature; E_m: auxiliary representation; r_m: contribution coefficient.

with modality features and performs routed fusion (Rao et al., 2026); and LSE-DGF adaptively balances original and generated descriptive information (Zhang et al., 2026).

However, auxiliary descriptions are not equally useful for every sample. Two issues can arise. First, descriptions from different modalities may provide inconsistent or unevenly informative interpretations of the same sample. Second, a generated description may not be fully supported by its own source signal. Therefore, the key question is not only how to generate auxiliary descriptions, but also how much each modality-specific auxiliary representation should contribute to the current prediction.

Existing methods already improve the use of auxiliary information through context filtering, feature–explanation alignment, routing, or gated fusion (Wang et al., 2024; Rao et al., 2026; Zhang et al., 2026). These approaches primarily select, align, or integrate auxiliary information; explicit pre-fusion contribution estimation for each modality-specific auxiliary representation remains less explored. Once auxiliary and source information have been jointly mixed, their individual effects become harder to regulate separately. Reliability-aware MSA addresses a related but different problem: CICA (Jiang et al., 2026a), 3WD-DRT (Jiang et al., 2026b), and other approaches (He et al., 2026; Zhu et al., 2026b; Zou et al., 2026) estimate the quality of original, degraded, or reconstructed modality representations. In contrast, the sample-specific contribution of LLM-generated auxiliary descriptions before source-feature integration remains comparatively underexplored.

To address this problem, we propose CONTRIFUSE, which adaptively regulates modality-specific auxiliary information before source-feature integration (Figure 1). CONTRIFUSE first constructs an auxiliary representation for each modality from descriptions generated by multiple LLMs. The resulting representations are then organized into shared and modality-specific components, allowing subsequent regulation to consider both information common across modalities and information specific to each modality. Based on the original modality feature, the organized auxiliary representation, cross-modal relations among auxiliary representations, and modality availability, Contribution-Aware Routing (CAR) estimates a sample- and modality-specific contribution coefficient. Multimodal Temporal Auxiliary Fusion (MTAF) then integrates the modulated auxiliary representation with the corresponding temporal modality feature for sentiment prediction.

Our main contributions are:

- We study how modality-specific auxiliary descriptions should contribute to multimodal sentiment prediction when their usefulness varies across samples and depends on both the source modality and other modality descriptions.

- We propose CONTRIFUSE, which organizes auxiliary representations into shared and modality-specific components and estimates their sample-specific contributions before source-feature integration using source features, cross-modal relations, and modality availability.
- We evaluate CONTRIFUSE on four English and Chinese MSA benchmarks using both regression and classification metrics, together with analyses of generator combinations, model components, routing inputs, modality availability, and contribution behavior.

2 RELATED WORK

2.1 MULTIMODAL REPRESENTATION LEARNING AND FUSION

Contextual models exploit neighboring utterances (Poria et al., 2017). TFN and LMF model high-order interactions (Zadeh et al., 2017; Liu et al., 2018); MulT captures cross-modal temporal dependencies (Tsai et al., 2019); and MISA, MMIM, Self-MM, and ALMT advance shared–private, mutual-information, unimodal-supervision, and hyper-modality learning (Hazarika et al., 2020; Han et al., 2021; Yu et al., 2021; Zhang et al., 2023). SentiLLM maps non-verbal signals to semantically aligned tokens for LLM reasoning (Chen et al., 2026).

2.2 AUXILIARY INFORMATION FOR MULTIMODAL SENTIMENT ANALYSIS

Recent methods make affective cues more explicit through auxiliary information. TextMI textualizes acoustic and visual cues for language-model reasoning (Hasan et al., 2023), DEVA generates emotional descriptions from audio and visual content (Wu et al., 2025), and WisdoM introduces contextual world knowledge with noise filtering (Wang et al., 2024). MSE-Adapter connects multimodal inputs to frozen LLMs (Yang et al., 2025b), TEXT aligns generated explanations with modality features and performs routed expert fusion (Rao et al., 2026), while LSE-DGF balances original and descriptively enhanced branches through difference-aware gating (Zhang et al., 2026). CONTRIFUSE estimates each modality-specific auxiliary representation's contribution before source-feature integration.

2.3 RELIABILITY AND SELECTIVE MULTIMODAL PROCESSING

Reliability-aware methods regulate multimodal information according to quality or confidence. CICA (Jiang et al., 2026a) uses confidence-aware pretraining and attention, 3WD-DRT (Jiang et al., 2026b) performs reliability-guided dynamic routing, EBMC balances modality collaboration, QA-MoE routes through quality-aware experts, and DEAR uses reconstruction fidelity under missing modalities (He et al., 2026; Zhu et al., 2026b; Zou et al., 2026). FRSR instead retrieves sentiment-relevant cues from original inputs before reasoning (Zhu et al., 2026a). These methods mainly regulate original, degraded, reconstructed, or retrieved modality information. CONTRIFUSE instead regulates the contribution of LLM-generated modality-specific auxiliary representations before source-feature integration, with the learned coefficient serving as a task-conditioned contribution signal.

3 METHOD

3.1 OVERALL FRAMEWORK

CONTRIFUSE combines temporal features with organized and contribution-regulated auxiliary representations (Figure 2). For sample n and modality $m \in \{\ell, a, v\}$ (text, audio, and vision), let $X_{n,m}^{\text{task}}$ denote the benchmark feature sequence and L_n its valid temporal length. A modality-specific feature encoder Enc_m^f produces

$$F_{n,m} = \text{Enc}_m^f(X_{n,m}^{\text{task}}) = [F_{n,m}^{(1)}, \ldots, F_{n,m}^{(L_n)}], \tag{1}$$

where $F_{n,m}$ is the encoded temporal feature sequence. Let T_n, A_n, and V_n denote the transcript, raw audio segment, and raw video segment, respectively. The modality-isolated textual inputs to the description generators are

$$\begin{aligned} Q_{n,\ell} &= T_n, \\ Q_{n,a} &= D_a(\text{eGeMAPS}(A_n)), \\ Q_{n,v} &= D_v(\text{OpenFace}(V_n)). \end{aligned} \tag{2}$$

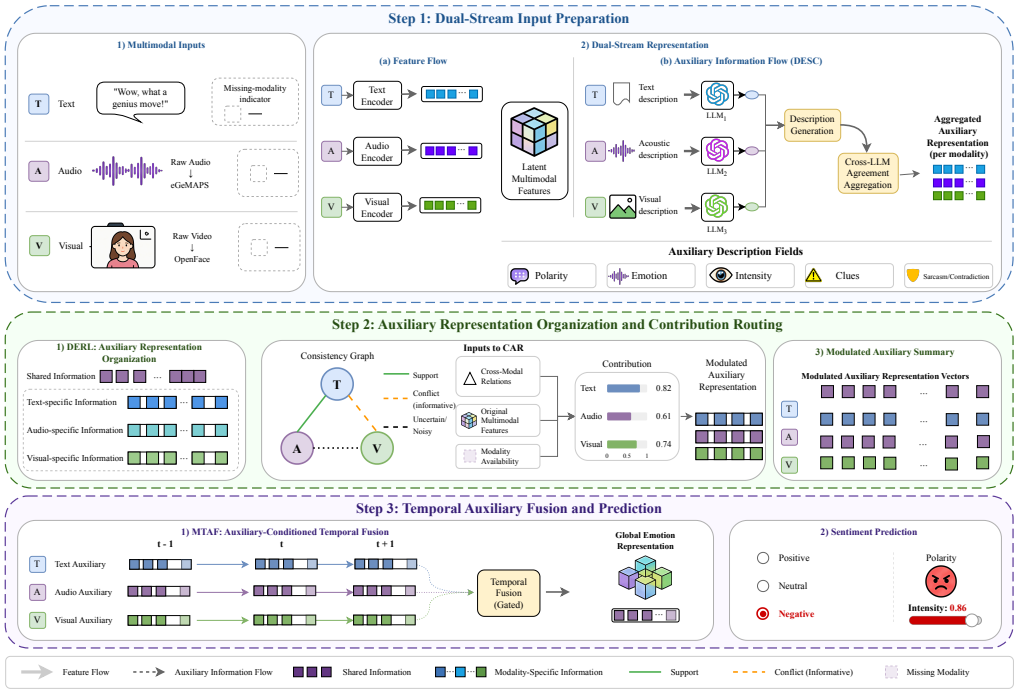


Figure 2: Overall framework. CONTRIFUSE retains the conventional temporal feature pathway and processes auxiliary descriptions in a parallel utterance-level pathway. Shared/modality-specific organization supports sample-specific pre-fusion contribution estimation; temporal modeling remains on the feature sequence.

D_a and D_v are deterministic textualization functions, not LLMs. Writing Q_n and F_n for the collections of modality inputs and features, the auxiliary-information path is

$$E_n = \mathrm{DESC}(Q_n), \qquad G_n = \mathrm{DERL}(E_n), \tag{3}$$

$$R_n = \mathrm{CAR}(G_n, F_n, M_n), \qquad \hat{y}_n = h(\mathrm{MTAF}(F_n, R_n, M_n)), \tag{4}$$

Here, E_n, G_n, and R_n denote the aggregated, organized, and contribution-modulated auxiliary representations, respectively; $M_n \in \{0,1\}^3$ denotes modality availability, and h is the sentiment regression head producing prediction $\hat{y}_n$. Diverse Description Stream Construction (DESC) aggregates descriptions; Disentangled Description Representation Learning (DERL) organizes their representations; CAR regulates their contributions; and MTAF integrates them with temporal features. The feature pathway preserves the source signal when auxiliary contribution is small.

3.2 DIVERSE DESCRIPTION STREAM CONSTRUCTION

DESC constructs modality-isolated auxiliary descriptions separately from downstream optimization; their representations can be cached for known samples. Audio descriptions use 88-dimensional eGeMAPS functionals (Eyben et al., 2016). Visual descriptions use OpenFace (Baltrušaitis et al., 2018) facial action units, valid-frame statistics, pose, and gaze. Normalization and discretization statistics are estimated only on the training split. The prompts receive only modality-isolated textual inputs $Q_{n,m}$: the transcript for text, an eGeMAPS-based acoustic description for audio, and an OpenFace-based visual description for vision. Raw audio and video, dataset and split identities, labels, scoring rules, and downstream predictions are not provided to the LLMs. For K description generators, generator $k \in \{1, \ldots, K\}$ produces

$$L_{n,m}^{(k)} = \mathrm{LLM}_k(Q_{n,m}), \qquad k \in \{1, \ldots, K\}, \tag{5}$$

where $L_{n,m}^{(k)}$ is the generated auxiliary description. Let Enc^e denote the frozen sentence encoder, P_{train} the PCA transformation fitted only on training-set description embeddings, and d_e the projected

auxiliary dimension. Both transformations remain fixed during downstream training:

$$Z_{n,m}^{(k)} = P_{\text{train}}\Big(\text{Enc}^e(L_{n,m}^{(k)})\Big) \in \mathbb{R}^{d_e}. \tag{6}$$

For $K \geq 2$, the cross-LLM agreement score $s_{n,m}^{(k)}$ and its softmax-normalized aggregation weight $\omega_{n,m}^{(k)}$ are

$$s_{n,m}^{(k)} = \frac{1}{K-1} \sum_{j \neq k} \text{Cos}(Z_{n,m}^{(k)}, Z_{n,m}^{(j)}), \tag{7}$$

$$\omega_{n,m}^{(k)} = \frac{\exp(s_{n,m}^{(k)})}{\sum_r \exp(s_{n,m}^{(r)})}, \qquad E_{n,m} = \sum_k \omega_{n,m}^{(k)} Z_{n,m}^{(k)}. \tag{8}$$

Here, Cos denotes cosine similarity. For $K = 1$, we bypass agreement and set $\omega_{n,m}^{(1)} = 1$ and $E_{n,m} = Z_{n,m}^{(1)}$. Cross-LLM agreement summarizes the semantic stability of independently generated interpretations and reduces dependence on any single generator. The weighted sum $E_{n,m}$ is the aggregated auxiliary representation for modality m, whose task contribution is subsequently estimated by CAR. Templates and generation settings appear in Appendix A.

3.3 Disentangled Description Representation Learning

Modality descriptions can share sentiment information while also providing modality-specific complementary cues. Treating all differences as noise would discard useful specificity, whereas modeling each modality independently would fail to capture cross-modal commonality. Building on the established shared–private representation paradigm (Hazarika et al., 2020), DERL uses a shared mapper D_s and a modality-specific mapper $D_{p,m}$ to obtain shared and modality-specific components:

$$S_{n,m} = D_s(E_{n,m}), \qquad P_{n,m} = D_{p,m}(E_{n,m}). \tag{9}$$

For a batch of N samples, shared components are encouraged to agree across available modalities,

$$\mathcal{L}_{\text{shared}} = \frac{1}{N} \sum_n \sum_{i<j} M_{n,i} M_{n,j} \|S_{n,i} - S_{n,j}\|_2^2, \tag{10}$$

while modality-specific components are encouraged to remain complementary,

$$\mathcal{L}_{\text{private}} = \frac{1}{N} \sum_n \sum_{i<j} M_{n,i} M_{n,j} \text{Cos}^2(P_{n,i}, P_{n,j}). \tag{11}$$

A modality-specific organization mapping ϕ_m^{org} combines the two components into the organized auxiliary representation:

$$G_{n,m} = \phi_m^{\text{org}}([S_{n,m}; P_{n,m}]). \tag{12}$$

Here, $[\cdot;\cdot]$ denotes concatenation. The shared component captures cross-modal commonality, while the modality-specific component preserves complementary information unique to each modality. The resulting representation $G_{n,m}$ therefore provides CAR with both shared and modality-specific information when estimating the contribution of each auxiliary representation.

3.4 Contribution-Aware Routing

Auxiliary descriptions can vary in relevance, consistency, and usefulness across samples and modalities, so their contribution should be determined separately for each sample and modality. The contribution of an auxiliary description depends on both its corresponding source modality and its relations with descriptions from other modalities. CAR therefore jointly considers the source-modality feature, the organized auxiliary representation, cross-modal relations among auxiliary representations, and modality availability to estimate a sample- and modality-specific contribution coefficient. Since $G_{n,m}$ is utterance-level and $F_{n,m}$ is temporal, valid source features are pooled:

$$\bar{F}_{n,m} = \text{Pool}(F_{n,m}^{(1)}, \ldots, F_{n,m}^{(L_n)}). \tag{13}$$

Here, $\mathrm{Pool}(\cdot)$ denotes validity-aware mean pooling. Pairwise relations are

$$c_{n,ij} = M_{n,i} M_{n,j} \mathrm{Cos}(G_{n,i}, G_{n,j}), \qquad c_{n,ii} = 0. \tag{14}$$

The four inputs have distinct roles. $\bar{F}_{n,m}$ provides the source-modality context; $G_{n,m}$ is the organized auxiliary representation; $c_{n,m}$ summarizes cross-modal semantic relations among auxiliary representations; and M_n states which modalities are available, changes how absent relations are interpreted, and supports hard masking. Using a modality-specific multilayer perceptron MLP_m and sigmoid activation $\sigma(\cdot)$, the coefficient is

$$r_{n,m} = M_{n,m}\, \sigma\big(\mathrm{MLP}_m[\bar{F}_{n,m}; G_{n,m}; c_{n,m}; M_n]\big)\,, \tag{15}$$

and the contribution-modulated auxiliary representation is

$$R_{n,m} = r_{n,m} G_{n,m}. \tag{16}$$

$r_{n,m}$ is learned end-to-end from sentiment supervision and serves as a task-conditioned contribution coefficient. If $M_{n,m} = 0$, the branch is hard-masked: its feature, auxiliary representation, pairwise relations, coefficient, and $R_{n,m}$ are zero. Pre-fusion regulation can therefore reduce the effect of ambiguous or inconsistent descriptions while preserving the corresponding original feature whenever that modality is available.

Regulating each auxiliary branch before multimodal fusion preserves explicit modality-specific contribution control prior to source-feature mixing.

3.5 Multimodal Temporal Auxiliary Fusion

MTAF integrates the contribution-modulated utterance-level auxiliary representation with the corresponding temporal feature sequence. The same $R_{n,m}$ conditions every valid feature step through the modality-specific feature–auxiliary conditioning mapping ϕ_m^{cond}:

$$\widetilde{F}_{n,m}^{(t)} = M_{n,m} \phi_m^{\mathrm{cond}}([F_{n,m}^{(t)}; R_{n,m}]), \qquad t = 1, \ldots, L_n. \tag{17}$$

The conditioning mapping ϕ_m^{cond} applies LayerNorm across the concatenated feature–auxiliary vector, followed by a linear projection, GELU activation, and dropout. This mapping is separate from ϕ_m^{org}. The conditioned modalities and availability are fused,

$$U_n^{(t)} = \phi_u([\widetilde{F}_{n,\ell}^{(t)}; \widetilde{F}_{n,a}^{(t)}; \widetilde{F}_{n,v}^{(t)}; M_n]), \tag{18}$$

where ϕ_u is the multimodal fusion mapping. Temporal dependencies are modeled by a Transformer followed by mask-conditioned hidden gating, attention pooling, and regression:

$$H_n = \mathrm{Transformer}(U_n^{(1)}, \ldots, U_n^{(L_n)}), \tag{19}$$

$$\widetilde{M}_n = \sigma(W_M M_n + b_M), \tag{20}$$

$$\widetilde{H}_n^{(t)} = \widetilde{M}_n \odot H_n^{(t)}, \tag{21}$$

$$\alpha_n^{(t)} = \frac{\exp(\boldsymbol{w}^\top \widetilde{H}_n^{(t)})}{\sum_i \exp(\boldsymbol{w}^\top \widetilde{H}_n^{(i)})}, \tag{22}$$

$$Z_n = \sum_t \alpha_n^{(t)} \widetilde{H}_n^{(t)}, \qquad \hat{y}_n = h(Z_n). \tag{23}$$

Here, H_n is the Transformer hidden sequence, $\widetilde{M}_n$ is the mask-conditioned hidden gate, $\widetilde{H}_n^{(t)}$ is the gated hidden state, $\alpha_n^{(t)}$ denotes temporal attention weights, and Z_n is the pooled utterance representation. W_M, b_M, and $\boldsymbol{w}$ are trainable parameters, $\odot$ is elementwise multiplication, and h is the regression head. Temporal modeling acts on the feature sequence; the auxiliary representation supplies a shared utterance-level condition across time. The objective is

$$\mathcal{L} = \mathcal{L}_{\mathrm{task}} + \lambda_s \mathcal{L}_{\mathrm{shared}} + \lambda_p \mathcal{L}_{\mathrm{private}}. \tag{24}$$

Here, $\mathcal{L}_{\mathrm{task}}$ is the MAE regression loss, and λ_s and λ_p weight the shared and modality-specific regularizers.

Description construction remains separate from downstream optimization, with frozen generators, sentence encoder, and train-only PCA. Implementation details appear in Appendix B.

Table 1: Dataset statistics.

Dataset	Train	Val.	Test	Total	Lang.
CMU-MOSI	1,284	229	686	2,199	EN
CMU-MOSEI	16,326	1,871	4,659	22,856	EN
CH-SIMS	1,368	456	457	2,281	ZH
CH-SIMSv2	2,722	647	1,034	4,403	ZH

4 EXPERIMENTS

4.1 DATASETS AND METRICS

We evaluate on CMU-MOSI, CMU-MOSEI, CH-SIMS, and CH-SIMSv2 using their official splits (Zadeh et al., 2016; 2018; Yu et al., 2020; Liu et al., 2022); Table 1 summarizes their sizes. English labels lie in $[-3, 3]$ and Chinese labels in $[-1, 1]$. Training uses sentiment-intensity regression and validation MAE for model selection. We report MAE, Pearson correlation, Acc-7/Acc-3, binary accuracy, and weighted F1, with classification metrics derived directly from the continuous regression predictions.

4.2 IMPLEMENTATION AND COMPARISON PROTOCOL

We compare feature-centric, reliability-aware, Chinese-MSA, and LLM-based methods, including the Chinese benchmarks KuDA, TETFN, and TMBL (Feng et al., 2024; Wang et al., 2023; Huang et al., 2024), together with DEVA and TEXT. Table 2 presents the comparison results on the four benchmarks. Experiments use PyTorch on an NVIDIA RTX 4090. Descriptions are generated by gpt-5.4-mini, deepseek-chat, and qwen-plus (OpenAI, 2026; DeepSeek-AI et al., 2024; Yang et al., 2025a), encoded, projected to 128 dimensions, and cached. We use AdamW, batch size 32, learning rate 10^{-4}, weight decay 10^{-5}, at most 40 epochs, and validation-MAE early stopping with patience 8; all main dimensions are 128. Implementation, caching, and sample-alignment details appear in Appendix D.

4.3 MAIN RESULTS

Table 2 shows consistently competitive performance across all four benchmarks. On MOSI and MOSEI, CONTRIFUSE obtains the best displayed MAE and correlation, while EBMC leads Acc-7 and MSE-LLaMA2-7B leads Has-0 Acc-2/F1 on MOSEI. On the Chinese benchmarks, CONTRIFUSE achieves higher correlation and classification scores than TEXT on CH-SIMS, although TEXT retains a lower MAE (0.353 vs. 0.361). On CH-SIMSv2, CONTRIFUSE obtains the best displayed results across all five metrics, including an MAE of 0.232 and a correlation of 0.858. Overall, these results demonstrate competitive performance across English and Chinese benchmarks and across regression and classification metrics. Table 3 further quantifies the contribution of CAR through controlled ablations.

4.4 ABLATION AND CONTRIBUTION ANALYSIS

Table 3 shows that single-generator descriptions outperform the feature-only model, multi-generator combinations are generally stronger, and the full pipeline performs best. Cross-LLM agreement provides a principled mechanism for aggregating information from multiple generators.

Removing DERL, CAR, or MTAF degrades performance on both datasets. CAR improves MAE from 0.617 to 0.602 on MOSI and from 0.486 to 0.472 on MOSEI over equal contribution. Removing cross-modal relation $c_{n,m}$ causes the largest MAE increase, while availability has the smallest effect under complete inputs because $M_n = [1, 1, 1]$ is constant.

Contribution coefficients vary across samples and modalities, as illustrated in Appendix C.11; Appendix C.13 compares pre- and post-fusion regulation. We additionally conduct a description-stratified evaluation to examine whether the benefit of contribution regulation varies with cross-modal agreement among auxiliary descriptions. The evaluation protocol and stratified results are presented in Appendix C.14.

Table 2: Main comparison on the English benchmarks CMU-MOSI and CMU-MOSEI and the Chinese benchmarks CH-SIMS and CH-SIMSv2. For MOSI/MOSEI, Acc-2 and F1 are reported under the Has-0/Non-0 settings. Best results are shown in bold. CONTRIFUSE results are averaged over five random seeds; standard deviations are reported in the "Std." row.

	CMU-MOSI					CMU-MOSEI				
Method	MAE↓	Corr↑	Acc-7↑	Acc-2↑	F1↑	MAE↓	Corr↑	Acc-7↑	Acc-2↑	F1↑
TFN	0.947	0.673	34.46	77.99/79.08	77.95/79.11	0.572	0.714	51.60	78.50/81.89	78.96/81.74
LMF	0.950	0.651	33.82	77.90/79.18	77.80/79.15	0.576	0.717	51.59	80.54/83.48	80.94/83.36
MulT	0.879	0.702	36.91	79.71/80.98	79.63/80.95	0.559	0.733	52.84	81.15/84.63	81.56/84.52
MISA	0.776	0.778	41.37	81.84/83.54	81.82/83.58	0.557	0.751	52.05	80.67/84.67	81.12/84.66
Self-MM	0.708	0.796	46.67	83.44/85.46	83.36/85.43	0.531	0.764	53.87	83.76/85.15	83.82/84.90
ALMT	0.712	0.793	46.79	83.97/85.82	84.05/85.86	0.530	0.774	53.62	81.54/85.99	81.05/86.05
3WD-DRT	0.701	0.794	48.42	83.41/86.51	85.28/87.08	0.518	0.779	53.83	84.19/87.33	83.48/88.26
CICA	0.630	0.855	49.56	88.19/90.24	88.14/90.23	0.489	0.856	55.29	84.72/90.18	85.15/90.16
EBMC	0.694	0.833	50.34	86.26/87.84	86.20/87.79	0.505	0.824	**57.32**	86.04/88.10	86.23/88.07
MSE-LLaMA2-7B	0.754	0.767	43.73	83.24/83.38	83.14/83.13	0.501	0.787	55.57	**86.74**/69.90	**86.51**/69.85
DEVA	0.730	0.787	46.32	84.40/86.29	84.48/86.30	0.541	0.769	52.26	83.26/86.13	82.93/86.21
TEXT	0.666	0.829	45.92	86.44/88.72	86.55/88.76	0.528	0.786	52.29	85.02/86.57	85.01/86.85
CONTRIFUSE	**0.602**	**0.871**	**51.02**	**89.05/91.18**	**89.02/91.12**	**0.472**	**0.873**	56.41	85.63/**91.05**	86.02/**91.00**
Std.	0.0075	0.007	0.62	0.93/0.87	0.91/0.85	0.0063	0.0040	0.19	0.25/0.22	0.28/0.24

	CH-SIMS					CH-SIMSv2				
Method	MAE↓	Corr↑	Acc-3↑	Acc-2↑	F1↑	MAE↓	Corr↑	Acc-3↑	Acc-2↑	F1↑
TFN	0.432	0.591	65.12	78.38	78.62	0.303	0.707	72.21	80.14	80.14
LMF	0.441	0.576	64.68	77.17	77.88	0.367	0.557	64.90	74.18	73.88
MulT	0.453	0.564	64.77	78.56	79.66	0.291	0.738	73.19	80.68	80.73
KuDA	0.408	0.613	66.52	80.74	80.71	0.271	0.759	76.21	82.11	82.04
TETFN	0.420	0.577	63.24	81.18	80.24	0.310	0.695	73.65	79.73	79.81
Self-MM	0.425	0.595	65.47	80.04	80.44	0.311	0.695	72.61	79.69	79.76
ALMT	0.408	0.594	65.86	78.77	78.71	0.308	0.700	71.86	79.59	79.51
3WD-DRT	0.408	0.627	66.64	81.37	81.91	0.276	0.764	75.86	83.05	84.01
CICA	0.378	0.754	76.37	86.00	85.68	0.245	0.842	80.56	85.98	85.89
TMBL	0.429	0.592	65.43	79.12	78.75	0.313	0.706	73.02	80.46	80.36
MSE-LLaMA2-7B	0.500	0.415	49.67	71.12	64.84	0.382	0.553	69.34	75.53	75.44
DEVA	0.424	0.583	65.42	79.64	80.32	0.278	0.781	76.82	83.64	83.51
TEXT	**0.353**	0.758	76.94	86.73	86.75	0.241	0.846	80.92	86.42	86.31
CONTRIFUSE	0.361	**0.772**	**77.82**	**87.21**	**86.94**	**0.232**	**0.858**	**81.74**	**87.10**	**86.96**
Std.	0.0048	0.0061	0.57	0.44	0.46	0.0039	0.0052	0.48	0.37	0.41

4.5 ROBUSTNESS TO MISSING MODALITIES

Missing modality is a robustness setting rather than the primary motivation. We evaluate seven availability patterns and compare CONTRIFUSE with MIT-FRNet (Li et al., 2024), CICA, and zero-shot gpt-5.4-mini in Figure 3. Unavailable branches are hard-masked throughout the feature, auxiliary, pairwise-relation, routing, and fusion computations. Table 4 further summarizes CONTRIFUSE across all seven availability settings, with Acc-2 and F1 reported under the Non-0 protocol.

On both datasets, retaining text keeps MAE close to the complete-input result. Across T, T+A, and T+V, MAE is 0.604–0.608 on MOSI and 0.473–0.488 on MOSEI, whereas removing text increases MAE to 1.377–1.383 and 0.749–0.770, respectively. This confirms the dominant role of text in these benchmarks while highlighting the more challenging nature of text-missing evaluation. Under these settings, CONTRIFUSE consistently achieves lower MAE than the compared methods for A+V, A-only, and V-only.

The relatively small gap between T-only and T+A/T+V suggests that text carries most of the sentiment signal, while audio and vision provide complementary refinements. The stronger performance of CONTRIFUSE under text-missing settings further demonstrates its robustness across different modality-availability patterns.

Table 3: Generator, component, and CAR-input analysis. G, D, and Q denote gpt-5.4-mini, deepseek-chat, and qwen-plus. Acc-2/F1 use Has-0.

Setting	MOSI MAE	Corr	Acc-2	F1	Acc-7	MOSEI MAE	Corr	Acc-2	F1	Acc-7
Feature only	0.642	0.842	87.21	87.18	45.86	0.505	0.842	83.74	84.26	52.10
DESC (G)	0.609	0.864	88.53	88.20	47.42	0.479	0.866	85.05	85.43	55.86
DESC (D)	0.613	0.861	88.31	88.02	47.13	0.482	0.863	84.91	85.21	55.62
DESC (Q)	0.617	0.857	87.94	87.70	46.98	0.485	0.860	84.62	84.98	55.35
DESC (G+D)	0.606	0.867	88.72	88.45	47.56	0.476	0.869	85.31	85.74	56.08
DESC (G+Q)	0.608	0.865	88.61	88.36	47.48	0.477	0.868	85.24	85.67	55.97
DESC (D+Q)	0.611	0.862	88.44	88.18	47.21	0.480	0.865	85.12	85.42	55.74
Full ContriFuse	**0.602**	**0.871**	**89.05**	**89.02**	**51.02**	**0.472**	**0.873**	**85.63**	**86.02**	**56.41**
w/o DERL	0.625	0.856	88.02	88.00	46.85	0.491	0.858	84.52	85.01	54.92
w/o CAR (equal contribution)	0.617	0.861	88.46	88.41	46.92	0.486	0.863	84.86	85.32	55.41
w/o MTAF	0.614	0.858	88.55	88.50	47.05	0.482	0.861	85.01	85.46	55.63
w/o source feature $\bar{F}_{n,m}$	0.611	0.865	88.70	88.65	49.30	0.479	0.867	85.17	85.58	55.84
w/o organized auxiliary representation $G_{n,m}$	0.609	0.867	88.81	88.74	49.72	0.478	0.868	85.24	85.65	55.96
w/o cross-modal relation $c_{n,m}$	0.613	0.864	88.62	88.56	48.91	0.481	0.865	85.03	85.44	55.67
w/o availability context M_n	0.604	0.870	88.96	88.90	50.61	0.474	0.872	85.52	85.90	56.22

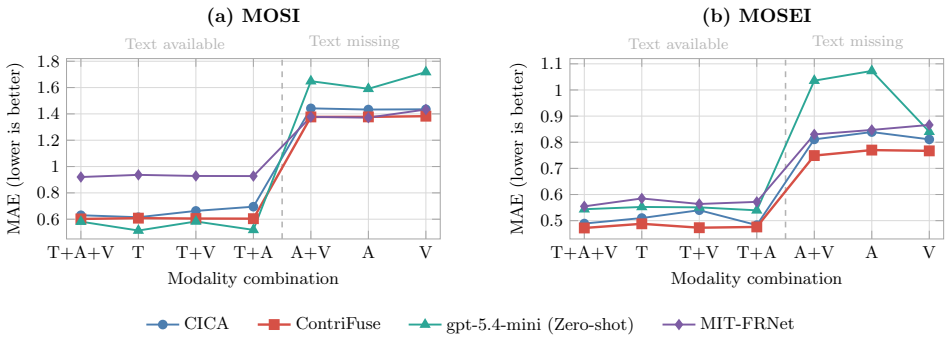


Figure 3: Missing-modality robustness on MOSI and MOSEI. Lower MAE is better. Missing modality is evaluated as a robustness setting rather than the primary motivation.

Table 4: Missing-modality robustness. Acc-2/F1 are Non-0 percentages.

Mod.	MOSI MAE	Corr	Acc-2	F1	MOSEI MAE	Corr	Acc-2	F1
T+A+V	**0.602**	**0.871**	**91.18**	**91.12**	**0.472**	**0.873**	**91.05**	**91.00**
T	0.608	0.866	90.69	90.58	0.488	0.857	90.90	90.82
T+V	0.605	0.870	90.33	90.23	0.473	0.871	90.22	90.17
T+A	0.604	0.868	90.43	90.36	0.476	0.868	90.23	90.14
A+V	1.377	0.249	51.78	36.21	0.749	0.469	74.39	73.32
A	1.377	0.289	59.02	43.85	0.770	0.399	71.95	68.08
V	1.383	0.287	59.19	44.90	0.767	0.402	70.61	69.20

5 Conclusion

ContriFuse addresses the unequal usefulness of modality-specific auxiliary descriptions by organizing shared and modality-specific information, estimating their sample-specific contribution before fusion, and integrating the modulated auxiliary representations with original temporal features. CAR uses source features, organized auxiliary representations, cross-modal relations, and modality availability to estimate task-conditioned contribution before fusion. Results on four English and Chinese benchmarks, including component, routing-input, and missing-modality analyses, support pre-fusion contribution regulation. The current auxiliary representation is utterance-level and depends on description quality; future work will study finer-grained and more efficient description generation.

AI USE DISCLOSURE

As part of DESC, we used gpt-5.4-mini, deepseek-chat, and qwen-plus to generate modality-specific auxiliary descriptions, as detailed in Section 3.2 and Appendix A. Generative AI tools assisted with language and minor code editing and with refining scientific figures. The authors developed the research ideas, methodology, experimental design, and conclusions; reviewed and revised the AI-assisted manuscript material; and take full responsibility for the final manuscript.

REFERENCES

Tadas Baltrušaitis, Amir Zadeh, Yao Chong Lim, and Louis-Philippe Morency. OpenFace 2.0: Facial behavior analysis toolkit. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pp. 59–66. IEEE, 2018. doi: 10.1109/FG.2018.00019.

Wei Chen, Junkai Li, Tongguan Wang, Hui Liu, Feiyue Xue, Chuanxiang Ma, and Ying Sha. Semantic-aligned structural abstraction for multimodal sentiment analysis. *arXiv preprint arXiv:2607.27790*, 2026.

DeepSeek-AI, Aixin Liu, Bei Feng, Bing Xue, et al. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*, 2024.

Florian Eyben, Klaus R. Scherer, Björn W. Schuller, Johan Sundberg, Elisabeth André, Carlos Busso, Laurence Y. Devillers, Julien Epps, Petri Laukka, Shrikanth S. Narayanan, and Khiet P. Truong. The geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing. *IEEE Transactions on Affective Computing*, 7(2):190–202, 2016. doi: 10.1109/TAFFC.2015.2457417.

Xinyu Feng, Yuming Lin, Lihua He, You Li, Liang Chang, and Ya Zhou. Knowledge-guided dynamic modality attention fusion framework for multimodal sentiment analysis. In *Findings of the Association for Computational Linguistics: EMNLP*, pp. 14755–14766, 2024.

Wei Han, Hui Chen, and Soujanya Poria. Improving multimodal fusion with hierarchical mutual information maximization for multimodal sentiment analysis. In *EMNLP*, 2021.

Md Kamrul Hasan, Md Saiful Islam, Sangwu Lee, Wasifur Rahman, Iftekhar Naim, Mohammed Ibrahim Khan, and Ehsan Hoque. Textmi: Textualize multimodal information for integrating non-verbal cues in pre-trained language models. *arXiv preprint arXiv:2303.15430*, 2023.

Devamanyu Hazarika, Roger Zimmermann, and Soujanya Poria. Misa: Modality-invariant and -specific representations for multimodal sentiment analysis. In *ACM Multimedia*, 2020.

Kang He, Yuzhe Ding, Xinrong Wang, Fei Li, Chong Teng, and Donghong Ji. Enhance-then-balance modality collaboration for robust multimodal sentiment analysis. In *CVPR*, 2026.

Jiehui Huang, Jun Zhou, Zhenchao Tang, Jiaying Lin, and Calvin Yu-Chian Chen. TMBL: Transformer-based multimodal binding learning model for multimodal sentiment analysis. *Knowledge-Based Systems*, 285:111346, 2024.

Alejandro Jaimes and Nicu Sebe. Multimodal human–computer interaction: A survey. *Computer Vision and Image Understanding*, 108(1–2):116–134, 2007.

Yanan Jia. A deep learning system for sentiment analysis of service calls. In *Proceedings of the 3rd Workshop on e-Commerce and NLP*, pp. 24–34, 2020.

Haoyu Jiang, Xiaoliang Chen, Duoqian Miao, Xiaolin Qin, Xianyong Li, and Yajun Du. Cica: Coupling confidence-aware pretraining with confidence-informed attention for robust multimodal sentiment analysis. In *CVPR*, 2026a.

Haoyu Jiang, Xiaoliang Chen, Duoqian Miao, Hongyun Zhang, Xiaolin Qin, Shangyi Du, and Peng Lu. 3wd-drt: A three-way decision enhanced dynamic routing transformer for cost-sensitive multimodal sentiment analysis. *Information Sciences*, 725:122704, 2026b.

Jiayao Li, Saihua Cai, Li Li, Ruizhi Sun, Gang Yuan, and Rui Zhu. Mit-frnet: Modality-invariant temporal representation learning-based feature reconstruction network for missing modalities. *Expert Systems with Applications*, 249:123655, 2024.

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 *ICMI*, 2022.

Zhun Liu, Ying Shen, Varun Bharadhwaj Lakshminarasimhan, Paul Pu Liang, Amir Zadeh, and Louis-Philippe Morency. Efficient low-rank multimodal fusion with modality-specific factors. In *ACL*, 2018.

OpenAI. Models: Openai api. `https://platform.openai.com/docs/models`, 2026. Accessed 2026-07-14.

Soujanya Poria, Erik Cambria, Devamanyu Hazarika, Navonil Majumder, Amir Zadeh, and Louis-Philippe Morency. Context-dependent sentiment analysis in user-generated videos. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, pp. 873–883, 2017.

Dongning Rao, Yunbiao Zeng, Zhihua Jiang, and Jujian Lv. A text-routed sparse mixture-of-experts model with explanation and temporal alignment for multi-modal sentiment analysis. In *AAAI*, 2026.

Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (eds.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 3982–3992, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1410. URL `https://aclanthology.org/D19-1410/`.

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 *ACL*, 2019.

Di Wang, Xutong Guo, Yumin Tian, Jinhui Liu, LiHuo He, and Xuemei Luo. TETFN: A text enhanced transformer fusion network for multimodal sentiment analysis. *Pattern Recognition*, 136: 109259, 2023.

Wenbin Wang, Liang Ding, Li Shen, Yong Luo, Han Hu, and Dacheng Tao. Wisdom: Improving multimodal sentiment analysis by fusing contextual world knowledge. In *ACM Multimedia*, 2024.

Sheng Wu, Dongxiao He, Xiaobao Wang, Longbiao Wang, and Jianwu Dang. Enriching multimodal sentiment analysis through textual emotional descriptions of visual-audio content. In *AAAI*, 2025.

An Yang et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025a.

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. In *AAAI*, pp. 25642–25650, 2025b. doi: 10.1609/aaai.v39i24.34755.

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 *ACL*, 2020.

Wenmeng Yu, Hua Xu, Ziqi Yuan, and Jiele Wu. Learning modality-specific representations with self-supervised multi-task learning for multimodal sentiment analysis. In *AAAI*, 2021.

Amir Zadeh, Rowan Zellers, Eli Pincus, and Louis-Philippe Morency. Multimodal sentiment intensity analysis in videos: Facial gestures and verbal messages. *IEEE Intelligent Systems*, 2016.

Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Tensor fusion network for multimodal sentiment analysis. In *EMNLP*, 2017.

Amir B. 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 *ACL*, 2018.

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 *EMNLP*, 2023.

Hengyuan Zhang, Aizihaierjiang Yusufu, Jiang Liu, Abidan Ainiwaer, and Aizierguli Yusufu. Image-text multimodal sentiment analysis with large model-generated descriptive semantics and difference-aware gated fusion. *Electronics*, 2026.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. BERTScore: Evaluating text generation with BERT. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=SkeHuCVFDr.

Yuan Zhang, Xiaomei Tao, Hanxu Ai, Tao Chen, and Yanling Gan. Multimodal emotion recognition by fusing video semantics in video learning scenarios. In *Neural Information Processing*, volume 15289 of *Lecture Notes in Computer Science*, pp. 1–15. Springer Nature Singapore, 2025. doi: 10.1007/978-981-96-6585-3_1. URL https://link.springer.com/chapter/10.1007/978-981-96-6585-3_1. ICONIP 2024.

Aoqiang Zhu, Min Hu, and Yan Xing. Fast retrieval and slow reasoning for explainable multimodal sentiment analysis. In *Findings of ACL*, 2026a.

Yitong Zhu, Yuxuan Jiang, Guanxuan Jiang, Bojing Hou, Peng Yuan Zhou, Ge Lin, and Yuyang Wang. QA-MoE: Towards a continuous reliability spectrum with quality-aware mixture of experts for robust multimodal sentiment analysis. In *ACL*, pp. 31689–31703, 2026b. doi: 10.18653/v1/2026.acl-long.1461.

Shihao Zou, Wei Wei, and Yongshuo Zhang. Dear: Distributional error-aware reliability for robust multimodal sentiment analysis with missing modalities. In *Findings of ACL*, 2026.

SUPPLEMENTARY MATERIAL FOR CONTRIFUSE: CONTRIBUTION-AWARE FUSION OF AUXILIARY DESCRIPTIONS FOR MULTIMODAL SENTIMENT ANALYSIS

Anonymous authors
Paper under double-blind review

A PROMPT TEMPLATES AND UTTERANCE-LEVEL AUXILIARY DESCRIPTION CONSTRUCTION

This appendix defines the notation and implementation used by Diverse Description Stream Construction (DESC). Auxiliary-description construction is executed separately from downstream optimization and can be precomputed and cached for known samples. Downstream training consumes benchmark feature arrays and the cached auxiliary-representation array. A previously unseen sample requires one auxiliary-description construction pass before prediction, whereas cached samples require no repeated LLM calls during downstream training or repeated inference.

A.1 NOTATION AND FEATURE–AUXILIARY REPRESENTATION

Let $n \in \{1, \ldots, N\}$ index samples, $m \in \{\ell, a, v\}$ index text, audio, and vision, $k \in \{1, \ldots, K\}$ index LLM description generators, and $t \in \{1, \ldots, L_n\}$ index aligned feature steps. The conventional input and encoded feature sequence are

$$X_{n,m}^{\text{task}} = \left[x_{n,m}^{(1)}, \ldots, x_{n,m}^{(L_n)}\right], \qquad F_{n,m} = \text{Enc}_m^f\left(X_{n,m}^{\text{task}}\right).$$

Here d_m is the input dimension of modality m, and $F_{n,m}^{(t)} \in \mathbb{R}^{d_f}$ is the encoded representation at step t. The released model uses $d_f = 128$.

The auxiliary-information pathway constructs a modality-specific input description $Q_{n,m}$ and obtains the k-th LLM-generated auxiliary description as

$$L_{n,m}^{(k)} = \text{LLM}_k\left(Q_{n,m}\right).$$

Each $L_{n,m}^{(k)}$ is utterance-level and has no independent time index. After encoding and aggregation, the cached utterance-level auxiliary representation is used to condition the temporally varying feature sequence.

A.2 MODALITY-SPECIFIC DESCRIPTION INPUTS

The code uses the following modality-isolated inputs:

$$Q_{n,\ell} = \mathcal{T}_n,$$

$$Q_{n,a} = D_a\left(\text{eGeMAPS}(\mathcal{A}_n); \Theta_a^{\text{train}}\right),$$

$$Q_{n,v} = D_v\left(\text{OpenFace}(\mathcal{V}_n); \Theta_v^{\text{train}}\right),$$

where $\mathcal{T}_n$, $\mathcal{A}_n$, and $\mathcal{V}_n$ are the transcript, raw audio segment, and raw video segment; D_a and D_v are deterministic textualization functions, not LLMs; and Θ_a^{train} and Θ_v^{train} contain statistics estimated from the training split only.

Table A1: Modality-specific auxiliary representation inputs implemented in the released code.

Modality	Source modality data	Input description Q	Granularity
Text	Transcript	Original transcript	Utterance
Audio	Raw audio	eGeMAPSv02 group description	Utterance
Vision	Raw video	OpenFace AU, validity, pose, and gaze description	Utterance

The downstream feature pathway does not receive the eGeMAPS or OpenFace descriptions. It continues to use the benchmark feature arrays $X_{n,m}^{\text{task}}$.

A.2.1 Text Input

The text description is the literal prefix `Original transcript:` followed by the stripped transcript. It excludes dataset identity, split information, labels, scoring rules, acoustic information, visual information, and downstream predictions.

A.2.2 Audio Description

openSMILE 2.5.0 extracts one 88-dimensional eGeMAPSv02 functional vector (Eyben et al., 2016) for each audio segment. The code groups standardized functionals into F0 level, F0 variation, loudness level, loudness variation, voicing, rhythm, spectral profile, and voice quality. For functional j,

$$z_{n,j} = \frac{A_{n,j} - \mu_j^{\text{train}}}{\widetilde{\sigma}_j^{\text{train}}}, \qquad \widetilde{\sigma}_j^{\text{train}} = \begin{cases} \sigma_j^{\text{train}}, & \sigma_j^{\text{train}} > 10^{-8}, \\ 1, & \text{otherwise.} \end{cases}$$

Within each acoustic group, the matching standardized values are averaged. The resulting group score is discretized using its training-set one-third and two-third quantiles. Thus, the code uses training-set tertiles, not fixed thresholds such as $\pm\gamma$.

The exact fixed-format fields are: F0 level, F0 variation, loudness level, loudness variation, voicing pattern, rhythmic pattern, spectral profile, and voice-quality profile.

A.2.3 Visual Description and Valid-Face Handling

OpenFace (Baltrušaitis et al., 2018) is invoked with action-unit, pose, and gaze outputs enabled. For frame t, define

$$b_n^{(t)} = \mathbb{I}\left(\text{success}_n^{(t)} > 0.5\right), \qquad \rho_n^{valid} = \frac{1}{L_n^{all}} \sum_{t=1}^{L_n^{all}} b_n^{(t)}.$$

If at least one frame satisfies $b_n^{(t)} = 1$, all visual statistics are computed only from those valid rows. If no frame is valid, the implementation falls back to the complete OpenFace table rather than dropping the sample or returning an empty description. The valid-frame ratio ρ_n^{valid} is retained as a separate visual descriptor.

The action-unit set is AU01, AU02, AU04, AU05, AU06, AU07, AU09, AU10, AU12, AU14, AU15, AU17, AU20, AU23, AU25, AU26, AU28, and AU45. For every available OpenFace intensity column

AUj_r, the code stores the population mean and population standard deviation over the selected rows. For every available activation column AUj_c, it stores the proportion of values greater than 0.5:

$$\mu_{n,j}^{AU} = \mathrm{mean}_t\left(u_{n,j}^{(t)}\right), \qquad \sigma_{n,j}^{AU} = \mathrm{std}_{t,\,ddof=0}\left(u_{n,j}^{(t)}\right),$$

$$a_{n,j}^{AU} = \mathrm{mean}_t\left[\mathbb{I}\left(c_{n,j}^{(t)} > 0.5\right)\right].$$

The visual input description verbalizes the mean intensity and activation ratio for each AU. The stored per-AU standard deviation is not verbalized as a separate field. Instead, the description includes an aggregate facial-action variation computed as the mean Euclidean norm of consecutive differences across all available AU intensity columns:

$$\Delta_n^{AU} = \mathrm{mean}_{t\geq 2} \parallel \mathbf{u}_n^{(t)} - \mathbf{u}_n^{(t-1)} \parallel_2 .$$

Head movement is computed in the same way from available pose rotation and translation columns, and gaze movement is computed from the two gaze-angle columns. The final visual description therefore contains valid-frame ratio, AU intensity levels, AU activation-frequency levels, facial-action variation, head movement, and gaze movement. All categorical levels are assigned by training-set tertiles.

A.3 Exact Prompt Construction

The term 'evidence' appears in the historical generation prompt; throughout the paper, the generated outputs are referred to as auxiliary descriptions. The released code sends one system message and one modality-specific user message. The system prompt is:

```
You extract modality-supported affective evidence. Use only the
supplied
single-modality input. Describe only the polarity, possible emotion
category,
and intensity supported by that modality. Do not output a dataset
label,
official sentiment class, regression score, dataset identity, or final
prediction.
Return strict JSON only.
```

The common user-message suffix is:

```
Return JSON only with these fields: polarity (string),
emotion_category (string), intensity (string), clues (list),
sarcasm_or_contradiction (string), reasoning (string).
Preserve uncertainty in the field values when the evidence is
ambiguous.
Single-modality input:
{DESCRIPTION}
```

It is preceded by `Input modality: text.`, `Input modality: audio.`, or `Input modality: vision.` and the following exact guidance:

```
Text: Analyze lexical meaning, affective phrases, negation, contrast,
ambiguity, sarcasm, and discourse cues.

Audio: Analyze combinations of pitch, loudness, rhythm, voice quality,
and
```

```
temporal acoustic patterns. Never map one descriptor directly to
emotion.

Vision: Analyze combinations of facial action units, activation
strength,
valid-frame evidence, temporal variation, gaze, and head pose. Never
map one
action unit directly to emotion.
```

The returned JSON object must contain exactly the six fields named above, and `clues` must be a list. Markdown-fenced JSON is accepted by the parser, but the stored object is reduced to the six required fields.

A.4 Multi-LLM Generation Configuration

Table A2: LLM description-generation configuration in the released code.

Configuration item	Code value
Generators	gpt-5.4-mini, deepseek-chat, qwen-plus
Temperature	0.0
Top-p	Omitted unless explicitly supplied
Maximum output tokens	1800
Generation seed	42 by default
Maximum attempts	5
Timeout	120 seconds per request
Backoff between failed attempts	2, 4, 8, and 16 seconds
Qwen option	enable_thinking=False
Output constraint	JSON object with six required fields

Each model receives the same modality-specific input description and schema, but generated auxiliary descriptions are produced independently and are not exposed to one another.

A.5 Description Encoding and Consistency Aggregation

The six-field JSON object is serialized with sorted keys and encoded using the frozen Sentence Transformers encoder (Reimers & Gurevych, 2019), `sentence-transformers/all-MiniLM-L6-v2`. The implementation applies attention-mask mean pooling:

$$u_{n,m}^{(k)} = \mathrm{Enc}_e\left(L_{n,m}^{(k)}\right) \in \mathbb{R}^{384}.$$

The reference encoding path fits one PCA using only the flattened training embeddings across samples, modalities, and LLMs, and then transforms validation and test embeddings without refitting:

$$Z_{n,m}^{(k)} = P_{train}\left(u_{n,m}^{(k)}\right) \in \mathbb{R}^{128}.$$

For each sample and modality with $K \geq 2$, cross-LLM consistency is

$$s_{n,m}^{(k)} = \frac{1}{K-1} \sum_{\substack{j=1 \\ j \neq k}}^{K} Cos\left(Z_{n,m}^{(k)}, Z_{n,m}^{(j)}\right).$$

For $K = 1$, the implementation bypasses agreement, sets $\omega_{n,m}^{(1)} = 1$, and returns $E_{n,m} = Z_{n,m}^{(1)}$. For $K \geq 2$, it uses a numerically stable softmax with implicit temperature $\tau = 1$:

$$\omega_{n,m}^{(k)} = \frac{\exp\left(s_{n,m}^{(k)}\right)}{\sum_{r=1}^{K} \exp\left(s_{n,m}^{(r)}\right)}, \qquad E_{n,m} = \sum_{k=1}^{K} \omega_{n,m}^{(k)} Z_{n,m}^{(k)}.$$

Cross-LLM consistency is computed only among generated auxiliary descriptions for the same sample and modality. It measures semantic stability and agreement, not factual correctness.

A.6 Auxiliary-Representation Cache Creation and Modality Availability

The auxiliary-description construction procedure requires a resolved audio and video path for every full-modality record that it processes; an unresolved manifest entry raises an error. It does not impute an unavailable raw modality from the remaining modalities. Auxiliary representations can be precomputed and cached for known samples, while a previously unseen sample requires one construction pass before prediction. The missing-modality experiments are downstream robustness evaluations: cached three-modality auxiliary representations are loaded first, and the availability vector acts as both routing context and a masking signal that determines which modality branches can contribute. DESC is not rerun for each masking pattern.

The reference description-generation JSON stores the sample ID, source modality, input description, model name, raw generation, and parsed JSON. The `api_base` field is not stored in these records. The reference encoder checks sample-ID equality across all modality-generator files and can write split-specific sample-ID arrays. The behavior of the supplied training cache and downstream loader is documented separately in Appendix D.4–D.5.

B Auxiliary Representation Organization, Contribution Routing, and Auxiliary-Conditioned Temporal Fusion

B.1 Feature and Auxiliary Streams

For each sample, the downstream model receives $F_{n,m}$, the cached aggregated auxiliary representation $E_{n,m}$, and a modality-availability mask

$$M_n = [M_{n,\ell}, M_{n,a}, M_{n,v}] \in \{0,1\}^3.$$

B.2 Disentangled Description Representation Learning

The code uses one shared mapper and three modality-specific mappers. The shared component captures cross-modal commonality, while the modality-specific component preserves complementary modality-dependent information:

$$S_{n,m} = D_s(E_{n,m}), \qquad P_{n,m} = D_{p,m}(E_{n,m}).$$

The shared and modality-specific regularizers are

$$\mathcal{L}_{shared} = \frac{1}{N} \sum_{n=1}^{N} \sum_{i<j} M_{n,i} M_{n,j} \parallel S_{n,i} - S_{n,j} \parallel_2^2,$$

$$\mathcal{L}_{private} = \frac{1}{N} \sum_{n=1}^{N} \sum_{i<j} M_{n,i} M_{n,j} Cos^2(P_{n,i}, P_{n,j}).$$

The organized auxiliary representation retains both components for downstream contribution estimation:

$$G_{n,m} = \phi_m^{\text{org}}([S_{n,m}; P_{n,m}]).$$

B.3 CONTRIBUTION-AWARE ROUTING

After feature encoding, valid temporal steps are mean-pooled:

$$\bar{F}_{n,m} = \frac{\sum_t q_{n,m}^{(t)} F_{n,m}^{(t)}}{\max\left(1, \sum_t q_{n,m}^{(t)}\right)},$$

where $q_{n,m}^{(t)}$ is the temporal validity mask after applying modality availability. Cross-modal relations are computed from organized auxiliary representations:

$$c_{n,ij} = M_{n,i} M_{n,j} \operatorname{Cos}(G_{n,i}, G_{n,j}), \qquad c_{n,ii} = 0.$$

For modality m, the complete row $c_{n,m} = [c_{n,m\ell}, c_{n,ma}, c_{n,mv}]$ is supplied to a modality-specific routing MLP together with the complete availability mask:

$$r_{n,m} = M_{n,m}\, \sigma\left(\mathrm{MLP}_m[\bar{F}_{n,m}; G_{n,m}; c_{n,m}; M_n]\right), \qquad R_{n,m} = r_{n,m} G_{n,m}.$$

The pooled source feature provides source-modality context, the organized auxiliary representation supplies the current auxiliary information, cross-modal relations describe the relations among modality-specific auxiliary representations, and the availability mask supplies the available-modality and hard-masking context. Here $r_{n,m}$ is a task-conditioned contribution coefficient rather than a factual-correctness probability, and $R_{n,m}$ is the resulting modulated auxiliary representation.

B.4 MULTIMODAL TEMPORAL AUXILIARY FUSION

The modulated utterance-level auxiliary representation is repeated at every aligned feature step:

$$F_{n,m}^{e(t)} = M_{n,m}\, \phi_m^{\mathrm{cond}}([F_{n,m}^{(t)}; R_{n,m}]), \qquad t = 1, \ldots, L_n.$$

The three conditioned branches and the complete availability mask are fused as

$$U_n^{(t)} = \phi_u([F_{n,\ell}^{e(t)}; F_{n,a}^{e(t)}; F_{n,v}^{e(t)}; M_n]).$$

A one-layer Transformer encoder models the fused sequence. The mask is also projected into the hidden dimension and applied after the Transformer:

$$H_n = \mathrm{Transformer}(U_n^{(1)}, \ldots, U_n^{(L_n)}),$$

$$\widetilde{M}_n = \sigma(W_M M_n + b_M), \qquad H_n^{e(t)} = \widetilde{M}_n \odot H_n^{(t)}.$$

Attention pooling and the regression head produce

$$Z_n = \sum_{t=1}^{L_n} \alpha_n^{(t)} H_n^{e(t)}, \qquad \widehat{y}_n = h(Z_n).$$

B.5 TRAINING OBJECTIVE

The implemented objective is

$$\mathcal{L} = \mathcal{L}_{task} + \lambda_s \mathcal{L}_{shared} + \lambda_p \mathcal{L}_{private}.$$

The default task loss is mean absolute error; mean squared error is the only alternative exposed by the training program. Binary and seven-class metrics are computed from the continuous regression

predictions after training. The released implementation does not use a separate classification head or cross-entropy objective.

B.6 Exact Missing-Modality Processing

The dataset object returns the normalized feature arrays and cached auxiliary representations together with M_n; it does not physically delete or pre-zero the source arrays. Inside the model, each feature sequence is encoded first and then multiplied by $M_{n,m}$. The auxiliary representation tensor is multiplied by $M_{n,m}$ at model entry before DERL. The temporal validity mask is computed from nonzero source-feature rows and intersected with modality availability. Consequently, unavailable modalities have zero encoded sequences, zero pooled features, zero auxiliary representations, zero shared and modality-specific components, zero organized auxiliary representations, a zero contribution coefficient, and zero conditioned MTAF branches, although their source arrays may still exist in the input batch.

Table B1: Missing-modality handling implemented by the downstream model.

Stage	Feature handling	Auxiliary handling	Mask use
Dataset	Normalize and return all arrays	Return cached row	Return M_n
Feature encoder	Encode, then zero unavailable branch	–	$M_{n,m}$
DERL	–	Zero before mapping	$M_{n,m}$ and pair mask
CAR	Zero pooled unavailable branch	Zero relation rows/columns and routing	Full M_n
MTAF	Zero unavailable conditioned branch	Modulated auxiliary representation only	Full M_n

B.7 Missing-Modality Sampling During Training

For training index i at epoch e, the code initializes a local pseudorandom generator with

$$s_{e,i} = s_0 + 1,000,003e + 9,176i.$$

With the default drop rate 0.6, $T + A + V$ is selected with probability 0.40; otherwise one of the six incomplete settings is sampled uniformly, giving probability 0.10 to each. Validation and standard test loaders use fixed $T + A + V$ availability. Seven-setting evaluation reuses one validation-selected checkpoint and applies each fixed availability vector to the same test split.

B.8 Training and Inference Boundary

DESC is executed separately from downstream optimization and receives no downstream gradient. The downstream model receives an auxiliary representation tensor of shape $[B, 3, d_e]$. For samples already represented in the cache, downstream training and repeated inference require no repeated LLM calls. A previously unseen sample requires one input-description construction, generation, encoding, projection, and aggregation pass before prediction.

C Reproducibility Scope of the Supplementary Material

C.1 Implemented and Reported Analyses

The released training program directly supports the full model and the `wo_desc`, `wo_derl`, `wo_car`, and `wo_mtaf` component settings. The experimental suite additionally includes four CAR input-

ablation variants that remove the source feature $\bar{F}_{n,m}$, organized auxiliary representation $G_{n,m}$, cross-modal relation $c_{n,m}$, or availability context M_n, respectively; all other architectural components and training protocols are kept unchanged. The evaluation program applies seven fixed modality combinations to one validation-selected checkpoint. Unit tests cover cross-LLM aggregation, cache reshaping, missing-modality sampling, loss masking, and the inability of altered unavailable inputs to change the prediction.

The four CAR input ablations in the main paper use the same experimental protocol. Appendices C.4–C.10 report additional computational-cost, auxiliary-representation dimension, cross-LLM, prompt, human-evaluation, and temporal-backbone analyses.

C.2 Metric Definitions

MOSI and MOSEI are trained as regression tasks. Acc-7 rounds predictions and labels after clipping them to $[-3, 3]$. Has-0 binary metrics classify values at zero using $\widehat{y} \geq 0$ and $y \geq 0$. Non-0 metrics first discard examples with $y = 0$ and then classify using strict positivity. The code reports weighted F1 for both binary variants.

For CH-SIMS and CH-SIMSv2, Acc-3 applies the same mapping to each continuous prediction and ground-truth label z:

$$c_3(z) = \begin{cases} \text{negative}, & z < -0.1, \\ \text{neutral}, & -0.1 \leq z \leq 0.1, \\ \text{positive}, & z > 0.1. \end{cases}$$

Acc-3 is the fraction of samples whose predicted and true classes match. The comparisons use the stored numerical values directly, without clipping or rounding. Acc-2 uses $c_2(z) = \mathbf{1}[z > 0]$: zero predictions and labels belong to the non-positive class, and all samples, including those in the three-class neutral interval, are retained. Chinese F1 uses these binary classes and the same ground-truth-support weighting as the English binary metrics: $\mathrm{F1}_w = \sum_{c \in \{0,1\}} (N_c/N)\mathrm{F1}_c$, where N_c is the number of true examples in class c and N is the total sample count. Accuracy and F1 are reported as percentages.

C.3 Interpretation Boundary

Cross-LLM consistency is a within-modality agreement statistic. CAR uses cross-modal relations computed from learned organized auxiliary representations. Neither quantity establishes factual correctness or causal attribution. Auxiliary description examples may be inspected as model inputs, but they should not be presented as human-validated findings unless the annotation protocol and raw ratings are released.

C.4 Scope of the Additional Diagnostics

The following tables report additional diagnostic experiments. Unless otherwise stated, performance analyses use MOSI and keep the remaining model and training protocol unchanged. gpt-5.4-mini, deepseek-chat, and qwen-plus denote the three description generators used in the corresponding generator analyses. Cross-LLM agreement denotes semantic stability rather than correctness, CAR coefficients denote task-conditioned contribution rather than factual-correctness probabilities, and MTAF applies the same modulated utterance-level auxiliary representation as a global condition at aligned feature steps.

C.5 Auxiliary Description Construction Cost

The MOSI statistics can be reproduced from the released minimal package. The MOSEI, CH-SIMS, and CH-SIMSv2 statistics are author-recorded generation logs from the full experimental runs and are reported for reference; the corresponding raw assets are not included in the released minimal package.

Table C1: Auxiliary-description construction scale. "Expected modality–generator records" denotes the number of sample–modality–generator combinations in the complete configuration, while "Successfully completed records" counts combinations for which a valid output was eventually obtained after the configured retry procedure. Retry attempts are not counted as separate records. Statistics for datasets other than MOSI are taken from the authors' full-run generation logs and are not directly reproducible from the released minimal package.

Dataset	Samples	Expected modality–generator records	Successfully completed records	Total time (h)
CMU-MOSI	2,199	19,791	19,791	2.86
CMU-MOSEI	22,856	205,704	205,704	28.74
CH-SIMS	2,281	20,529	20,529	3.05
CH-SIMSv2	4,403	39,627	39,627	5.79

The completed-record count should not be interpreted as the number of first-attempt API successes. Individual requests could fail or time out and be retried according to the procedure in Table A2; the table reports only whether each sample–modality–generator task ultimately produced a valid cached output.

These measurements characterize one-time auxiliary-description construction for known samples. Each complete-modality sample produces at most nine structured auxiliary descriptions, one for each modality–generator pair; the table therefore reports sample–modality–generator record counts rather than API attempt counts or tokens per individual call. Total time is dataset-level wall-clock time and should not be interpreted as isolated single-call latency because calls may be scheduled in batches and include provider-side variation and retry handling. A cached sample requires no repeated LLM calls during downstream training or repeated inference, whereas a previously unseen sample still requires one auxiliary-description construction pass before prediction.

C.6 Projected Auxiliary-Representation Dimension

We further examine the effect of the projected auxiliary-representation dimension d_e on CMU-MOSI. The frozen `all-MiniLM-L6-v2` encoder produces a 384-dimensional sentence embedding for each generated auxiliary description. For dimensionality-reduction settings, a PCA transformation is fitted using training embeddings only and then applied unchanged to the validation and test splits. The 384-dimensional setting uses the original sentence embedding without PCA.

Table C2: Projected auxiliary-representation dimension analysis on MOSI.

Dimension	MAE ↓	Corr ↑	Acc-2 ↑	F1 ↑
384 (no PCA)	0.604	0.869	88.86	88.80
256	0.603	0.870	88.94	88.89
128	**0.602**	**0.871**	**89.05**	**89.02**
64	0.611	0.861	88.11	87.86

The 128-dimensional projection gives the strongest overall result while retaining a compact representation. Reducing the auxiliary representation dimension further to 64 causes a clearer degradation. The final reference path therefore uses 128-dimensional projected auxiliary representations, as documented in Appendices A.5 and D.4–D.6.

C.7 Cross-LLM Semantic Agreement

Table C3: Cross-LLM semantic agreement among gpt-5.4-mini, deepseek-chat, and qwen-plus, measured by cosine similarity and BERTScore (Zhang et al., 2020).

Generator pair	Cosine ↑	BERTScore ↑
gpt-5.4-mini vs. deepseek-chat	0.852	0.786
gpt-5.4-mini vs. qwen-plus	0.778	0.712
deepseek-chat vs. qwen-plus	0.784	0.739
Average	0.805	0.746

The values indicate substantial, but not perfect, semantic agreement among the generators. They describe stability across generated auxiliary descriptions and do not establish that the descriptions are factually correct.

C.8 Prompt Robustness

Table C4: Prompt-robustness analysis on MOSI.

Prompt	MAE ↓	Corr ↑	Acc-2 ↑	F1 ↑
Prompt A (default)	**0.602**	**0.871**	**89.05**	**89.02**
Prompt B	0.606	0.868	88.83	88.79
Prompt C	0.605	0.869	88.90	88.86
Average	0.604	0.869	88.93	88.89
Std.	0.002	0.001	0.092	0.096

Std. is the population standard deviation of the three reported prompt scores; Acc-2 and F1 deviations are in percentage points. Prompt A is the default prompt used for the complete-model result reported in the main paper. Prompts B and C are wording variants evaluated with the same description schema, data split, downstream architecture, and Has-0 metric definitions. Their Acc-2 scores are 0.22 and 0.15 percentage points below the default, respectively, while the three-prompt average is 0.12 points below it. The small but nonzero decrease therefore reflects limited sensitivity to the tested rephrasings rather than a change in evaluation protocol.

C.9 Human Evaluation of Generated Descriptions

Table C5: Human ratings for 100 generated-description samples.

Metric	Score (1–5) ↑
Faithfulness	4.38 ± 0.55
Reasonableness	4.35 ± 0.57
Informativeness	4.47 ± 0.49
Average	4.40 ± 0.54

Three independent annotators rated 100 randomly selected samples on five-point Likert scales. These ratings indicate that the generated auxiliary descriptions were generally judged faithful, reasonable, and informative, but they do not establish causal faithfulness of the final prediction.

C.10 TEMPORAL-BACKBONE ANALYSIS

Table C6: Temporal-backbone comparison on MOSI.

Backbone	MAE ↓	Corr ↑	Acc-2 ↑	F1 ↑
BiLSTM	0.610	0.864	88.62	88.31
GRU	0.612	0.862	88.40	88.18
Transformer	0.602	0.871	89.05	89.02

The Transformer performs best, although the gaps are modest. In every setting, the temporal backbone operates on the aligned feature sequence, with the same modulated utterance-level auxiliary representation serving as a global condition.

C.11 REPRESENTATIVE CONTRIBUTION CASES

To provide an intuitive view of how contribution-aware routing (CAR) regulates auxiliary information at the sample and modality levels, we examine representative samples from the MOSI test set. Table C7 presents modality-specific auxiliary descriptions together with the learned contribution coefficients r_T, r_A, and r_V. All case-study records were matched by sample ID to the MOSI test split before extracting contribution coefficients and auxiliary descriptions.

Complementary. For video `tStelxIAHjw`, segment 10, the label is -0.4 and the prediction is -0.6823. The text auxiliary description has a neutral majority polarity (2/3), with cues "didn't think; was sad." The audio description is negative (2/3), with low F0 level, limited F0 variation, high loudness level, and strong loudness variation. The visual description is also negative (3/3), with strong AU04, weak AU06/AU12, low facial variation, and limited head movement. The coefficients are $r_T = 0.273$, $r_A = 0.811$, and $r_V = 0.597$. Here, the non-text auxiliary descriptions provide additional negative affective information that complements the weaker, near-neutral sentiment expressed in the text description.

Consistent. For video `f9O3YtZ2VfI`, segment 21, the label is 2.0 and the prediction is 1.9877. All three auxiliary descriptions have a positive majority polarity, and their coefficients are relatively balanced: $r_T = 0.553$, $r_A = 0.580$, and $r_V = 0.604$. This example illustrates that CAR can preserve comparable contributions when modality-specific auxiliary descriptions convey mutually consistent sentiment information.

Inconsistent / weak support. For video `jUzDDGyPkXU`, segment 2, the label is 2.6 and the prediction is 2.3014. The text and audio descriptions are positive (both 3/3), whereas the visual description is mixed (2/3). The visual coefficient, $r_V = 0.268$, is substantially lower than $r_T = 0.636$ and $r_A = 0.756$. The mixed visual description differs from the consistently positive text and audio descriptions and receives a smaller contribution, consistent with CAR's design for sample-specific contribution regulation.

These cases are qualitative illustrations and do not establish that contribution coefficients measure factual correctness. The coefficients are learned through sentiment supervision and represent task-conditioned contribution. In particular, the mixed visual description does not by itself establish an error or a lack of support from the source signal.

Table C7: Representative MOSI test cases illustrating modality-specific auxiliary descriptions and learned contribution coefficients. The descriptions summarize the polarity agreed upon by the three LLMs together with representative modality cues.

Case	Text auxiliary description	Audio auxiliary description	Visual auxiliary description	r_T	r_A	r_V
Complementary	*Consensus: neutral (2/3); cues: didn't think; was sad.*	*Consensus: negative (2/3); cues: low F0 level; limited F0 variation; high loudness level; strong loudness variation.*	*Consensus: negative (3/3); cues: strong AU04; weak AU06 and AU12; low facial variation; limited head movement.*	0.273	0.811	0.597
Consistent	*Consensus: positive (3/3); cues: my two favorite; Tilda Swinton; Cate Blanchett.*	*Consensus: positive (2/3); cues: high loudness level; strong loudness variation; fast speech rate; high voiced proportion.*	*Consensus: positive (3/3); cues: moderate AU06; weak AU12; weak rare AU04; low facial variation.*	0.553	0.580	0.604
Inconsistent / weak support	*Consensus: positive (3/3); cues: really really well directed; the action.*	*Consensus: positive (3/3); cues: wide F0 variation; high loudness; strong loudness variation; high voiced segment proportion.*	*Consensus: mixed (2/3); cues: AU04 strong, rare; AU06 weak, rare; AU12 strong, intermittent; high facial-action variation.*	0.636	0.756	0.268

Counts (2/3 or 3/3) indicate agreement on polarity among the three LLMs; representative cues are retained from the exported descriptions.

C.12 Failure-Case Analysis

Two representative errors illustrate limitations that remain after task-conditioned routing. For the utterance *Yeah, that was just great.*, the ground-truth label is negative whereas the model predicts neutral. The surface wording is positive and the available generated auxiliary descriptions do not resolve the sarcastic interpretation strongly enough. This case emphasizes that cross-LLM semantic agreement is not a correctness guarantee: agreement among weak interpretations cannot supply a cue that is absent from the generated auxiliary descriptions.

For the utterance *I think the movie was okay.*, the ground-truth label is neutral whereas the model predicts negative. Most auxiliary descriptions support a neutral interpretation, but the visual stream contains a misleading negative cue. CAR regulates the task-conditioned contribution of that modality-specific auxiliary representation, yet a nonzero routed contribution can still influence the final feature–auxiliary fusion. These cases motivate explicit uncertainty modeling and counterfactual contribution analysis rather than interpreting CAR as a correctness assessor.

C.13 Pre-Fusion versus Post-Fusion Contribution Regulation

To examine whether the location of contribution regulation matters, we compare the pre-fusion design used in CONTRIFUSE with a controlled post-fusion variant. In the full model, the learned coefficient is applied directly to the organized auxiliary representation before source-feature integration:

$$R_{n,m} = r_{n,m} G_{n,m}. \tag{25}$$

For the post-fusion variant, we retain the same router architecture and input types but apply contribution regulation after feature–auxiliary integration. Specifically, we first compute

$$\begin{aligned} H_{n,m}^{(t),\mathrm{full}} &= \phi_m^{\mathrm{cond}}([F_{n,m}^{(t)}; G_{n,m}]), \\ H_{n,m}^{(t),\mathrm{feat}} &= \phi_m^{\mathrm{cond}}([F_{n,m}^{(t)}; 0]), \end{aligned} \tag{26}$$

and then interpolate the two representations using

$$\widetilde{H}_{n,m}^{(t)} = (1 - r_{n,m}) H_{n,m}^{(t),\mathrm{feat}} + r_{n,m} H_{n,m}^{(t),\mathrm{full}}. \tag{27}$$

Pre-fusion regulation scales the auxiliary input before joint LayerNorm and GELU-based nonlinear transformation. In contrast, the post-fusion variant interpolates between feature-only and feature–auxiliary outputs computed by the same parameter-shared conditioner. For an affine conditioning map, the two formulations coincide; under the implemented nonlinear conditioner, they are generally different. This comparison therefore evaluates input-space regulation versus output-space interpolation within the same conditioning architecture.

Table C8: Comparison of equal contribution, post-fusion regulation, and the proposed pre-fusion CAR on MOSI and MOSEI. Lower MAE and higher correlation are better.

	MOSI		MOSEI	
Setting	MAE ↓	Corr ↑	MAE ↓	Corr ↑
Equal contribution	0.617	0.861	0.486	0.863
Post-fusion regulation	0.615	0.866	0.483	0.867
CAR (pre-fusion)	**0.602**	**0.871**	**0.472**	**0.873**

Post-fusion regulation improves MAE and correlation over equal contribution on both datasets. Pre-fusion CAR further improves both metrics, reducing MAE from 0.615 to 0.602 on MOSI and from 0.483 to 0.472 on MOSEI. This result supports regulating each modality-specific auxiliary representation before it is mixed with the source feature, where its individual contribution can be controlled more directly.

C.14 Contribution Regulation under Cross-Modal Description Disagreement

This analysis examines whether the benefit of contribution regulation depends on agreement among modality-specific auxiliary descriptions. We compare full CONTRIFUSE with equal contribution and the variant without cross-modal relation inputs on CMU-MOSI under complete-modality inputs. The comparison holds the source features, cached auxiliary representations, data splits, and training protocol fixed across variants. Each variant is evaluated using its own validation-selected checkpoint, with sample IDs matched before computing paired prediction errors.

For each modality, a majority polarity category requires agreement from at least two of the three description generators. Samples with no majority category in at least one modality are assigned to the no-majority group first. The remaining samples are partitioned into three mutually exclusive groups: matching majority categories across all three modalities, opposite-polarity disagreement containing both positive and negative categories, and other disagreement. The mixed category is retained as a distinct category rather than converted to positive or negative. Group assignments depend solely on the original generated polarity fields and are fixed across model variants and evaluation seeds. Ground-truth labels, downstream predictions, contribution coefficients, and learned relation scores are excluded from the grouping procedure. The four groups contain 52 matching-category samples, 261 opposite-polarity samples, 266 other-disagreement samples, and 107 no-majority samples.

We report MAE for each subgroup, including samples with zero sentiment labels. For subgroup $\mathcal{S}_g$ and comparison variant b, the paired MAE reduction is defined as

$$\Delta_g^{(b)} = \frac{1}{|\mathcal{S}_g|} \sum_{i \in \mathcal{S}_g} \left(|y_i - \hat{y}_i^{(b)}| - |y_i - \hat{y}_i^{\mathrm{Full}}| \right),$$

where positive values indicate lower prediction error for full CONTRIFUSE. The main diagnostic is whether these reductions are larger for opposite-polarity descriptions than for matching majority categories.

Table C9: MOSI performance stratified by cross-modal description agreement under complete-modality inputs. The Overall row corresponds to Table 3. Lower MAE is better.

Description group	N	Equal contribution	w/o cross-modal relation	Full CONTRIFUSE
Matching majority categories	52	0.500	0.499	0.498
Opposite-polarity disagreement	261	0.638	0.630	0.610
Other disagreement	266	0.642	0.640	0.634
No majority in at least one modality	107	0.560	0.560	0.553
Overall	**686**	**0.617**	**0.613**	**0.602**

Table C9 shows that adaptive contribution regulation provides the largest MAE reductions when auxiliary descriptions express opposing sentiment interpretations. On the opposite-polarity subset, CONTRIFUSE obtains an MAE of 0.610, compared with 0.638 for equal contribution and 0.630 for the variant without cross-modal relation inputs. The corresponding MAE reductions are 0.028 and 0.020, whereas the reductions in the matching-category group are 0.002 and 0.001, respectively.

The reductions are smaller in the other-disagreement group, at 0.008 and 0.006, and are 0.007 against both comparison variants in the no-majority group. This pattern indicates a larger benefit on opposite-polarity samples. Weighting the subgroup MAEs by sample count yields the overall values of 0.617, 0.613, and 0.602 after rounding.

The larger paired reductions on opposite-polarity samples support the usefulness of sample-specific contribution regulation when modality-specific auxiliary descriptions conflict. The additional advantage over the no-relation variant is consistent with cross-modal relation inputs helping contribution estimation under these conditions.

D Implementation Details and Reproducibility Checks

D.1 Interpretable Non-Text Feature Configuration

Audio input descriptions use openSMILE eGeMAPSv02 88-dimensional functionals. Visual input descriptions use OpenFace AU intensity and activation columns together with face-detection success, pose, and gaze columns. These descriptors are used by auxiliary-description construction, which is executed separately from downstream optimization; the downstream feature pathway uses the benchmark feature pickle.

Table D1: Implemented non-text description sources.

Modality	Tool	Source input	Description content
Audio	openSMILE 2.5.0	Raw audio	Eight training-normalized eGeMAPS groups
Vision	OpenFace	Raw video or OpenFace CSV	Validity, 18 AUs, variation, pose, gaze

D.2 eGeMAPS Group Construction

The eight group matchers operate on eGeMAPSv02 column names: F0 level uses F0 semitone means; F0 variation uses F0 standard-deviation and percentile-range terms; loudness level uses loudness means; loudness variation uses loudness standard-deviation and percentile-range terms; voicing uses voiced, unvoiced, and voicing-probability terms; rhythm uses segments-per-second and segment-length terms; spectral profile uses spectral flux, alpha ratio, Hammarberg index, and MFCC terms; and voice quality uses jitter, shimmer, HNR, and harmonic-difference terms.

D.3 OpenFace Statistics

The visual-statistics computation follows Appendix A.2.3. The implementation accepts an existing OpenFace CSV or runs the configured OpenFace binary on raw video. Valid-frame selection, the fallback when no valid frame exists, all 18 AU names, and the pose/gaze trajectory summaries are fixed in the released descriptor code.

D.4 Auxiliary-Representation Cache Storage Layout

The reported results use the reference encoding path. It applies one unified train-only PCA and writes split arrays with shape $[N, 3, 128]$, together with projected auxiliary representations, cross-LLM consistency scores, aggregation weights, and split sample IDs. The cached representation therefore matches the unified projection P_{train} defined in the main paper rather than a modality-specific PCA variant. The downstream loader checks the documented shape, auxiliary representation dimension, finite values, and split alignment before model use.

D.5 Sample-ID Handling During Downstream Training

The generation and encoding path used for the reported results records sample IDs and checks equality across the nine modality-generator files before aggregation. Split-specific sample-ID arrays are retained with the cache and checked against the feature-split order before downstream training. Thus, row i in each auxiliary-representation split corresponds to row i in the associated feature split; an ID or row-count mismatch fails validation rather than being silently reordered.

D.6 Projection Fitting and Data Isolation

All 384-dimensional training embeddings are flattened across samples, modalities, and LLMs to fit one unified PCA. Validation and test embeddings are transformed with the stored training mean and components and are never used to refit the projection. The PCA and sentence encoder are not updated during downstream training.

D.7 Description Generation Configuration

The generation settings are listed in Table A2. Existing JSON caches are accepted only when their stored schema equals the six required description fields. Completed records are skipped on restart. Each newly generated record stores the raw and parsed response immediately after a successful call.

D.8 Missing-Modality Leakage Checks

The model-flow tests compare predictions after replacing unavailable feature and auxiliary branches with large random values. Because the model applies the availability mask after feature encoding and at auxiliary-representation entry, the prediction must remain unchanged. Separate tests verify pairwise loss masking and the configured modality-sampling distribution.

D.9 Missing-Modality Training Distribution

Table D2: Default training distribution over modality availability.

Availability setting	Sampling probability
$T + A + V$	0.40
$T + A$	0.10
$T + V$	0.10
$A + V$	0.10
T	0.10
A	0.10
V	0.10

D.10 Main Training Configuration

Table D3: Downstream training defaults in the released code.

Configuration item	Code default
Optimizer	AdamW
Initial learning rate	1×10^{-4}
Weight decay	1×10^{-5}
Batch size	32
Maximum epochs	40
Early-stopping patience	8 epochs
Validation criterion	MAE
Gradient clipping	1.0
Task loss	MAE
λ_s, λ_p	0.05, 0.02
Feature, organized, and fusion dimensions	128
Dropout	0.2
Transformer heads and layers	4 heads, 1 layer
Training drop rate	0.6
Data-loader workers	2
Runtime device	auto: CUDA when available, otherwise CPU

D.11 Cache and Dataset Validation

The supplied MOSI feature pickle is checked for the expected train, validation, and test counts; required feature and regression-label keys; finite feature values; and the recorded SHA-256 checksum. The auxiliary-representation cache used for the reported results is checked for split keys, shape $[N, 3, 128]$, finite values, unified-PCA metadata, auxiliary representation dimension, and split sample IDs. Row counts and sample-ID order must agree with the corresponding feature splits.

D.12 Summary

The released code reproduces the downstream CONTRIFUSE data flow, component switches, three-term loss, deterministic missing-modality sampling, seven fixed evaluation settings, and metric definitions. The exact prompt, visual statistics, invalid-frame fallback, cache layout, and sample-alignment assumptions are stated above. Additional diagnostic experiments are reported in Appendices C.4–C.14.

E Limitations and Scope of Use

Although CONTRIFUSE achieves strong performance on multiple multimodal sentiment analysis benchmarks and is additionally evaluated under missing-modality conditions, several limitations remain. We summarize them in terms of generated description reliability, feature interpretation, temporal granularity, computational cost, interpretability, privacy, and practical scope of use.

E.1 Generated Description Reliability and LLM Dependence

CONTRIFUSE relies on multiple LLMs to generate modality-specific auxiliary descriptions from transcripts, eGeMAPS-based acoustic descriptions, and OpenFace-based visual descriptions. Cross-LLM consistency aggregation can reduce the influence of isolated abnormal outputs, while DERL and CAR further organize and regulate the resulting auxiliary representations. However, these mechanisms do not guarantee that generated descriptions are always correct.

Multiple generators may produce similar but biased interpretations, especially for ambiguous, sarcastic, context-dependent, or culturally dependent samples. Moreover, semantic agreement among

generators reflects stability rather than factual correctness. A low-consistency description may occasionally capture a useful minority interpretation that other generators miss, while high consistency may also arise from shared model biases or similar pretraining distributions.

Because the pretraining corpora of the external LLMs are not fully disclosed, we also cannot independently verify whether public benchmark utterances or related web content appeared during pretraining. To reduce direct evaluation leakage, prompts exclude dataset identity, data splits, labels, label definitions, scoring rules, target regression scores, and downstream predictions. Nevertheless, complete absence of pretraining overlap cannot be guaranteed. Future work may explicitly model generator uncertainty and disagreement, preserve multiple candidate interpretations, and reduce dependence on external LLMs through local or distilled description generators.

E.2 Feature Interpretation and Cross-Cultural Generalization

The acoustic and visual descriptions used by CONTRIFUSE are constructed from interpretable measurements such as eGeMAPS descriptors and OpenFace facial action units. These features provide traceable observations for auxiliary description generation, but individual measurements do not deterministically correspond to specific emotions. For example, similar pitch, loudness, facial action, gaze, or head-pose patterns may reflect different affective states depending on speaker characteristics, recording conditions, interaction context, or social behavior.

The current evaluation includes both English and Chinese benchmarks, but this does not establish universal cross-lingual or cross-cultural generalization. The affective meaning of lexical, prosodic, and facial cues may vary across languages, cultures, individuals, and contexts. CONTRIFUSE therefore treats these signals as informative observations rather than direct emotion labels. Future work may study multilingual auxiliary-description encoders, cross-cultural human evaluation, and language- or culture-conditioned description generation.

E.3 Temporal Granularity

DESC generates one utterance-level aggregated auxiliary representation for each sample and modality. The LLM is not invoked independently for every word, acoustic frame, visual frame, or local event. This design reduces generation cost by avoiding dense frame-level description generation, but it also limits the ability to represent fine-grained temporal changes.

A long utterance may contain local shifts in attitude, sarcasm, emphasis, or facial behavior that are compressed into one global auxiliary representation. MTAF mitigates this limitation by using the modulated utterance-level auxiliary representation as a global condition over the aligned temporal feature sequence, but this is not equivalent to constructing time-varying auxiliary descriptions. Future work may explore clause-level textual descriptions, sliding-window acoustic descriptions, action-unit-event visual descriptions, and hierarchical reasoning across local and utterance-level representations.

E.4 Computational Cost

For samples whose auxiliary representations have already been constructed and cached, downstream training and repeated inference require no additional LLM calls. However, a previously unseen full-modality sample may require up to nine description-generation calls in the current three-modality, three-generator configuration. It also requires acoustic and visual descriptor extraction, description construction, sentence encoding, projection, and cross-LLM aggregation.

Therefore, the current framework is most practical when auxiliary representations can be constructed once and reused. Future work may reduce this cost by invoking multiple generators only for uncertain samples, using a single generator for easier cases, dynamically selecting generators according to modality informativeness, distilling multi-LLM descriptions into lightweight local models, or replacing external APIs with local foundation models.

E.5 Interpretability and Scope of Use

The structured auxiliary descriptions, cross-LLM aggregation, DERL organization, and CAR contribution coefficients improve the analyzability of CONTRIFUSE. However, readable auxiliary descriptions should not be interpreted as complete causal explanations of the final prediction. The final output also depends on conventional multimodal features, representation mappings, cross-modal relations, contribution coefficients, temporal fusion, Transformer encoding, attention, and the prediction head. In particular, CAR coefficients represent task-conditioned contribution rather than probabilities of factual correctness.

The experimental conclusions of this work are limited to multimodal sentiment analysis benchmarks. Model outputs are statistical predictions under specific data and annotation protocols rather than objective measurements of an individual's psychological state. CONTRIFUSE should therefore not be directly used for high-risk decisions such as clinical diagnosis, mental-health risk assessment, hiring or promotion, disciplinary decisions, insurance or credit decisions, criminal-risk prediction, immigration screening, or judgments of individual trustworthiness. Applications beyond benchmark sentiment analysis require task-specific validation, appropriate human oversight, and evaluation under the target deployment setting.

E.6 Privacy and Third-Party Models

Although the current pipeline does not directly send raw audio or video files to external LLMs, transcripts, eGeMAPS-based acoustic descriptions, OpenFace facial-action descriptions, and associated metadata may still contain sensitive or identifying information. Feature descriptions should therefore not be treated as inherently anonymous.

When applying CONTRIFUSE to real user data, deployment should consider data licensing, informed consent, de-identification, data minimization, access control, retention policies, and the data-processing policies of third-party model providers. Local feature processing and private or locally deployed language models may be preferable when privacy requirements are strict.