CaReFlow: Cyclic Adaptive Rectified Flow for Multimodal Fusion (Appendix)

Supplementary Material

1. Unimodal Networks

In this section, we introduce the architectures of unimodal networks and detail the processes for generating unimodal representations that will be fused later on. To be consistent with state-of-the-art models [12, 32], we employ pre-trained language models [5, 14, 16] to obtain high-level language features. Specifically, for all the downstream tasks, the procedures of the language network are presented as below:

$$\begin{aligned} \hat{\boldsymbol{X}}_l &= \text{PLM}(\boldsymbol{U}_l;\ \theta_l) \in \mathbb{R}^{T\times d_l} \\ \boldsymbol{X}_l &= (\hat{\boldsymbol{X}}_l \boldsymbol{W}_{pro} + \boldsymbol{b}_{pro}) \in \mathbb{R}^{T\times d} \end{aligned} \tag{1}$$

where PLM denotes the pre-trained language model, $\boldsymbol{U}_l$ is the input token sequence and T is the sequence length. $\boldsymbol{W}_{pro} \in \mathbb{R}^{d_l\times d}$ and $\boldsymbol{b}_{pro} \in \mathbb{R}^{1\times d}$ are trainable parameters that map the output dimensionality of the language network to the shared dimensionality d. For the MSA task, the procedures for the acoustic and visual networks, which incorporate transformer encoders [29], are shown as follows ($m \in \{a, v\}$):

$$\begin{aligned} \hat{\boldsymbol{X}}_m &= \text{Conv 1D}(\boldsymbol{U}_m;\ K_m) \in \mathbb{R}^{T\times d} \\ \boldsymbol{X}_m &= \text{Transformer}(\hat{\boldsymbol{X}}_m;\ \theta_m) \in \mathbb{R}^{T\times d} \end{aligned} \tag{2}$$

where Conv 1D represents the temporal convolution whose kernel size K_m is set to 3. The generated unimodal representation $\boldsymbol{X}_m$ is used for distribution mapping.

To ensure effective extraction of humor-related features for the MHD and MSD tasks, and in alignment with prior methods [12, 22], we additionally extract a humor-centric feature (HCF) from the language modality. This serves as the fourth modality (see [12] for details) and is denoted as $\boldsymbol{U_h} \in \mathbb{R}^{T\times d_h}$. Furthermore, for MHD and MSD, each multimodal sample is composed of a target punchline segment and its preceding context segment. We concatenate the punchline and context feature sequences in the time dimension to obtain the final unimodal inputs $\boldsymbol{U_m} \in \mathbb{R}^{T\times d_m}$ ($m \in \mathcal{M} = \{a, v, l, h\}$). The unimodal network for the HCF modality, which also employs transformer encoders, follows a similar architecture to those used for the visual and acoustic modalities. Specifically, the transformer-based unimodal network operations are defined as follows (for $m \in \{a, v, h\}$):

$$\begin{aligned} \hat{\boldsymbol{X}}_m &= \text{Transformer}(\boldsymbol{U}_m;\ \theta_m) \in \mathbb{R}^{T\times d_m} \\ \boldsymbol{X}_m &= \text{Conv 1D}(\hat{\boldsymbol{X}}_m;\ K_m) \in \mathbb{R}^{T\times d} \end{aligned} \tag{3}$$

To reduce the model complexity of the subsequent processing, we fuse the language and HCF modalities via a simple linear layer:

$$\boldsymbol{X}_l \longleftarrow \text{Linear}(\boldsymbol{X}_l \oplus \boldsymbol{X}_h;\ \theta_{lin}) \in \mathbb{R}^{T\times d} \tag{4}$$

2. Datasets

(1) **CMU-MOSI** [35]: The CMU-MOSI dataset, a popular dataset for multimodal sentiment analysis (MSA), consists of more than 2,000 video segments sourced from the Internet. For each segment, sentiment intensity is annotated on a Likert scale ranging from -3 to 3, with 3 indicating the most intense positive sentiment and -3 indicating the most intense negative sentiment.

(2) **CMU-MOSEI** [36]: The CMU-MOSEI dataset is a widely-used large scale multimodal dataset for MSA collected from the Internet, encompassing over 22,000 video segments from over 1,000 YouTube speakers on over 250 varied topics. These segments are picked at random from a wide variety of topics and solo video presentations. Each segment is marked with two sets of annotations: emotions divided into six specific categories and sentiment scores that extend from -3 to 3. To evaluate CaReFlow on the MSA task, we utilize sentiment labels from the CMU-MOSEI dataset, which mirror the sentiment scale of the CMU-MOSI dataset.

(3) **CH-SIMS-v2** [18]: The CH-SIMS-v2 dataset is a Chinese MSA dataset collected from 11 distinct scenarios including interviews, talk shows, and films to simulate real-world human-computer interactions. The videos were filtered to ensure high-quality acoustic and visual features. The dataset is partitioned into training, validation, and test sets in a 9:2:3 ratio. To be more specific, the training set contains 2,722 video segments, further categorized into 921 negative, 433 weakly negative, 232 neutral, 318 weakly positive, and 818 positive samples. The validation and test sets are composed of 647 and 1,034 video segments respectively.

(4) **UR-FUNNY** [11]: The UR-FUNNY dataset is derived from TED talk videos with 1,741 speakers for the multimodal humor detection (MHD) task. Each target video segment in the dataset is called punchline, which contains language, acoustic, and visual modalities. The segments preceding the punchline are the context segments, which are fed into the model together with the punchline for contextual analysis. The punchlines are identified using the 'laughter' tag in the transcripts, which indicates when the audience laughed during the talk. Negative samples are similarly identified, where the target punchline segments are not followed by the 'laughter' tag. UR-FUNNY is divided into a training set with 7,614 samples, a validation set with

980 samples, and a testing set with 994 samples. In line with state-of-the-art methods [12, 20, 22], we employ version 2 of UR-FUNNY for our experimental analysis.

(5) **MUStARD** [2]: The MUStARD dataset, designed for detecting sarcasm in multimedia data, is derived from well-known TV shows including Friends, The Big Bang Theory, The Golden Girls, and Sarcasmaholics. It contains 690 video segments that are manually categorized as sarcastic or non-sarcastic. Except from the punchline segments that are the focus, MUStARD incorporates the preceding conversations (context segments) to provide contextual clues.

3. Evaluation Metrics

For the CMU-MOSI and CMU-MOSEI datasets, we evaluate the performance of CaReFlow and baselines using the following evaluation metrics: (1) **Acc7**: Evaluates the model's capability to classify sentiment scores into seven distinct categories; (2) **Acc2**: Assesses the model's ability to differentiate positive from negative sentiments in a binary classification scenario; (3) **F1 score**: The harmonic mean of precision and recall for binary sentiment classification (predictions are rounded to the nearest integer between -3 and 3 for Acc7; neutral segments are excluded when computing Acc2 and F1 score); (4) **MAE**: The mean absolute error between model predictions and annotated labels; and (5) **Corr**: The correlation coefficient indicating the strength and direction of the relationship between the model's predictions and the annotators' assessments. Notably, to compute Acc7, we round up the prediction to an integer from -3 to 3. For the CH-SIMS-v2 dataset, we use Acc5, Acc3, Acc2, F1 score, MAE, and Corr to evaluate the performance of the model.

For MHD and MSD, in alignment with prior methodologies [12, 20, 22], we report the binary accuracy (i.e., humorous or non-humorous, sarcastic or non-sarcastic) of the model.

3.1. Feature Extraction Details

(1) **Visual Modality**: For the CMU-MOSI and CMU-MOSEI datasets, following previous approaches [23, 32], Facet [1] is used to gather an array of visual attributes, such as facial action units, facial landmarks, and head positioning. These attributes are extracted for each segment, creating a time-based series of visual features that illustrate the progression of facial expressions and body gestures. For the CH-SIMS-v2 dataset, to be consistent with previous works [8, 18], OpenFace [1] is used to extract facial features such as 68 facial landmarks, 17 facial action units, head pose, head orientation, and eye gaze direction. For the UR-Funny and MUStARD datasets (MHD and MSD tasks), in line with prior approaches [12, 20], we also use OpenFace 2 [1] to extract facial action unit features as well as rigid and non-rigid facial shape parameters.

[1] iMotions 2017. https://imotions.com/

Table 1. The feature dimensionality of various datasets.

	Language	Acoustic	Visual	HCF
CMU-MOSI	768	74	47	-
CMU-MOSEI	768	74	35	-
CH-SIMS-v2	768	25	177	-
UR-FUNNY	768	60	36	4
MUStARD	768	60	36	4

(2) **Acoustic Modality**: For the CMU-MOSI, CMU-MOSEI, UR-Funny, and MUStARD datasets, COVAREP [4] is used to extract a sequence of acoustic features. This set of features includes 12 Mel-frequency cepstral coefficients, pitch tracking, speech polarity, glottal closure instants, and spectral envelope, among other elements. These features form a sequence that captures the dynamic changes in vocal tone throughout the speech. For the CH-SIMS-v2 dataset, 25-dimensional eGeMAPS low-level descriptors (LLD) features are extracted using the OpenSmile [6] at a sampling rate of 16000 Hz.

(3) **Language Modality**: For the CMU-MOSI and CMU-MOSEI datasets, following state-of-the-art methods [32], DeBERTa [14] is employed to extract textual features. For the CH-SIMS-v2 dataset, following previous works [8, 18], BERT [5] is used to extract textual features. For the MHD and MSD tasks, following state-of-the-art methods [12, 20, 22], ALBERT [16] is applied as the language network. Notably, for MHD and MSD, we concatenate the punchline and context token sequences to create the final input for the language network: $\boldsymbol{U_l} = \boldsymbol{C_l} \oplus [SEP] \oplus \boldsymbol{P_l}$, where the $[SEP]$ token is used to separate the context tokens $\boldsymbol{C_l}$ from the punchline tokens $\boldsymbol{P_l}$ [12].

Please refer to Table 1 for the input feature dimensionality of different modalities. For detailed information on the feature extraction process for the HCF modality, please see [12].

4. Experimental Details

We implement our CaReFlow using the PyTorch framework on an NVIDIA RTX3090 GPU with CUDA version 11.4 and PyTorch version 1.13.1. The training of CaReFlow is facilitated by the AdamW [19] optimizer. Please refer to Table 2 for detailed information on the hyperparameter settings. Following previous works [9, 24], we conduct a random grid search with fifty random iterations to search for optimal hyperparameters using the validation set, and save the hyperparameter setting that reaches the best performance. After the best hyperparameter setting is determined, we train the model again with the best hyperparam-

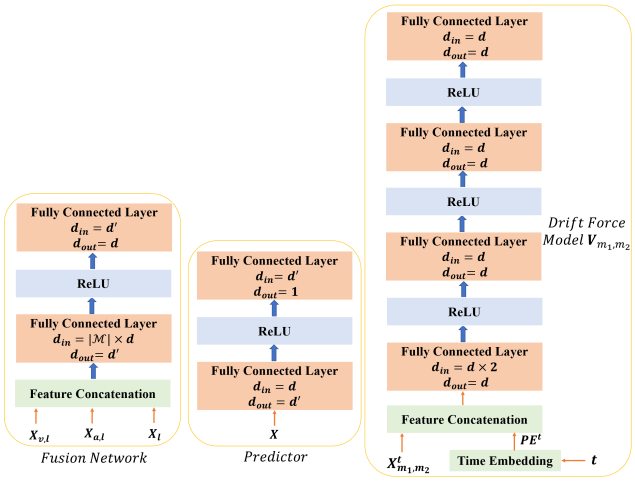

Figure 1. **The structures of multimodal fusion network, predictor, and drift force model V_{m_1,m_2}.** d' represents the hidden dimensionality, and $|\mathcal{M}|$ is the number of modalities.

eter setting for five times, and the final results are obtained by calculating the mean results of the five-time running.

The structures of the multimodal fusion network, predictor, and drift force model $\boldsymbol{V}_{m_1,m_2}$ are shown in Figure 1. Notably, CaReFlow scales efficiently to N modalities by extending pairwise mappings (e.g., MulT) to central anchor mapping, which only trains $N-1$ forward and $N-1$ backward models by defining anchor modality. Each forward/backward model is a lightweight MLP (72K parameters on MOSI), almost negligible compared to total parameters (185M).

Notably, we use 'one-to-many' and 'one-to-one' mappings in the forward flow and backward flow, respectively. We do not use one rectified flow model to realize both forward flow and backward flow because: (1) Forward flow uses 'one-to-many' mapping for global distribution learning, while backward flow uses 'one-to-one' mapping (same-sample pairs) to ensure point-wise reconstruction of source features. A single flow cannot handle both. (2) Backward flow enforces transformed features retain source information to regularize forward model. Separation prevents excessive interference with distribution-mapping learning.

In the 'one-to-many' mapping of forward flow, it is common to transform a source point to a target point that is from a different class. When transforming $\boldsymbol{X}_{m_1}$ (happy) to $\boldsymbol{X}_{m_2}$ (angry), the goal is not to 'convert happy to angry' but to learn distributional structure of target modality. $\boldsymbol{X}_{m_2}$ (angry) is only used to inform $\boldsymbol{X}_{m_1}$ what the language distribution space is. In other words, CaReFlow learns how to map a visual feature of 'happy' to language distribution space across all sentiments, i.e., where it should locate in language space. Moreover, by applying relaxed alignment, CaReFlow knows a 'happy' visual feature should be closer to the word 'happy' in language space, slightly farther from 'angry', and much farther from spaces of irrelevant modalities.

5. Baselines

The baselines for MSA include:

(1) **DEVA** [31]: It generates textual sentiment descriptions from audio-visual inputs to enhance emotional cues, and employs a text-guided progressive fusion module to improve alignment and fusion in nuanced emotional scenarios; (2) **Complete Multimodal Information Bottleneck** (**C-MIB**) [21]: It utilizes the information bottleneck principle to reduce redundancy and noise within unimodal and multimodal features, serving as the baseline for handling noisy modalities; (3) **Information-Theoretic Hierarchical Perception** (**ITHP**) [32]: Based on the information bottleneck principle, ITHP defines a core modality and regards the remaining modalities as detectors within the information pathway that serve to distill the flow of information; (4) **Multimodal Boosting** [24]: It integrates multiple base learners within a boosting-like framework, where each learner targets distinct aspects of multimodal learning. To assess the contribution of individual learners, Multimodal Boosting introduces a contribution learning module that dynamically determines both the contribution and noise level of each base learner; (5) **Self-Supervised Multi-task Multimodal sentiment analysis network** (**Self-MM**) [34]: It generates sentiment labels for individual modalities using the global labels of multimodal samples in a self-supervised manner, which helps extract more discriminative unimodal features; (6) **Disentangled-Language-Focused Model** (**DLF**) [30]: It proposes a feature disentanglement module to separate modality-shared and modality-specific information, and four geometric measures are introduced to refine the disentanglement process, reducing redundancy and enhance language-targeted features. It also designs a language-focused attractor to strengthen language representation by leveraging complementary modality-specific information; (7) **Attention-based Causality-Aware Fusion** (**AtCAF**) [15]: AtCAF leverages a counterfactual cross-modal attention module to identify causal relationships within the training data, establishing a comprehensive causal chain that effectively maps the causal path from user inputs to model outputs; (8) **Decoupled Multimodal Distillation** (**DMD**) [17]: To enhance the discriminative features of each modality, DMD introduces a flexible and adaptive cross-modal knowledge distillation approach. It decouples each unimodal representation into modality-irrelevant and modality-exclusive subspaces, and then employs a graph distillation unit to process each decoupled component in a more targeted and effective manner; (9) **Enhanced Dynamic Emotion Experts** (**EMOE**) [7]: It consists of mixture of modality experts for dy-

Table 2. Hyperparameter Settings of CaReFlow. MAE, MSE and BCE denote mean absolute error, mean square error and binary cross-entropy, respectively.

	CMU-MOSI	CMU-MOSEI	CH-SIMS-v2	MUStARD	UR-FUNNY
Loss Function	MSE	MSE	MAE	BCE	BCE
Batch Size	48	64	32	64	40
Learning Rate	1e-5	1e-5	5e-4	4e-6	3e-6
Sample Ratio β	4	1	4	1	2
Number of Euler Steps $\frac{1}{dt}$	2	2	2	1	2
Forward Loss Weight α_f	0.1	0.005	0.01	0.01	0.05
Backward Loss Weight α_b	0.1	0.005	0.01	0.01	0.001
Shared Dimensionality d	120	200	100	100	150
Hidden Dimensionality $d^{'}$	128	64	100	100	150
Margin ϵ	1e-4	0.001	1e-5	5e-5	1e-6

namically adjusting modality importance based on sample features, and unimodal distillation to retain unimodal predictive ability within fused features; (10) **Multimodal Adaptation Gate BERT/ALBERT** (**MAG-BERT/MAG-ALBERT**) [26]: It integrates the visual and acoustic information into BERT/ALBERT by introducing a multimodal adaptation gate; (11) **Acoustic Visual Mix-up Consistent** (**AV-MC**) [18]: It employs modality mix-up to enhance visual and acoustic representations, strengthening the role of visual and acoustic modalities in sentiment analysis; (12) **Knowledge-Guided Dynamic Modality Attention Fusion** (**KUDA**) [8]: It guides multimodal fusion with emotional knowledge by dynamically selecting the dominant modality and adjusting the contributions of each modality; (13) **MISA** [13]: It decomposes unimodal features into modality-invariant and modality-specific features, and then fuses them to generate prediction; (14) **MultiModal InfoMax** (**MMIM**) [10]: It enhances multimodal representation by maximizing mutual information between unimodal features as well as between different-level multimodal representations and unimodal features.

Notably, we implement the variants of C-MIB, Multimodal Boosting, ITHP, and DLF that use the same DeBERTa-v3 backbone as our CaReFlow for the CMU-MOSI and CMU-MOSEI datasets.

The additional baselines for MHD and MSD include:

(1) **Multimodal Global Contrastive Learning** (**MGCL**) [22]: MGCL implements supervised contrastive learning on multimodal representations and designs various strategies to generate positive and negative samples for each representation; (2) **Multimodal Correlation Learning** (**MCL**) [20]: MCL formulates a supervised multimodal correlation learning task that maintains modality-specific information while achieving a more discriminative embedding space; (3) **Modality Order-driven module for Sarcasm detection** (**MO-Sarcation**) [27]: It introduces a modality order-driven module incorporated into a transformer network to enable the ordered fusion of modalities; (4) **Multimodal Multitask Interaction Learning** (**MIL**) [37]: It is a framework for joint detection of sarcasm and sentiment that consists of a cross-modal target attention mechanism to automatically learn the alignment between texts and images/speeches and a multimodal interaction learning paradigm to simultaneously capture the commonness and uniqueness of sarcasm and sentiment; (5) **Subject Causal Intervention** (**SuCI**) [33]: It introduces a simple yet effective causal intervention module that isolates the influence of subjects acting as unobserved confounders, enabling unbiased predictions by leveraging true causal effects; (6) **Humor Knowledge Enriched Transformer** (**HKT**) [12]: HKT offers an effective approach for MHD and MSD by leveraging humor-centric features as external knowledge to capture the ambiguity and sentiment information embedded within the language modality; (7) **Multimodal Transformer** (**MulT**) [28]: MulT employs multiple cross-modal transformers to translate source modalities to target modalities, thereby reduce the gap between modalities; (8) **Multimodal Learning using Optimal Transport** (**MuLOT**) [25]: MuLOT uses self-attention to exploit intra-modal correspondence and optimal transport for cross-modal correspondence. It uses multimodal attention fusion to capture the inter-dependencies between modalities.

6. Additional Experiments

Table 3. Comparison with simple transformation baselines.

Model	MOSI (Acc7/Acc2/MAE)	MOSEI (Acc7/Acc2/MAE)
Linear Layer	43.8 / 86.3 / 0.674	53.7 / 87.3 / 0.530
MLP (4 Layers)	45.3 / 87.3 / 0.666	54.0 / 86.7 / 0.523
MLP + Contrastive Learning	47.3 / 87.5 / 0.643	53.1 / 87.1 / 0.522
CaReFlow	**50.6 / 89.8 / 0.616**	**55.7 / 87.9 / 0.504**

Table 4. Discussion on anchor (target) modality on CH-SIMS-v2.

Anchor	Acc5↑	Acc3↑	Acc2↑	F1↑	MAE↓	Corr↑
Language Anchor	**57.9**	**75.9**	**82.9**	**82.9**	0.277	**0.745**
Acoustic Anchor	56.3	75.6	82.1	82.2	**0.270**	0.743
Visual Anchor	57.7	75.4	80.4	80.4	0.271	0.739

6.1. Discussion on Target Modality

We choose language as the target modality, and transform the distributions of visual and acoustic modalities to language distributions. This is because: (1) **Affect Dominance**: Language directly conveys sentiment with explicit semantic information, while visual/audio modalities provide indirect and complementary cues; and (2) **Distribution Consistency**: Language features such as BERT embeddings have more consistent distributional properties compared to visual/acoustic features that are sensitive to pose, lighting, etc. Aligning to a consistent language distribution simplifies cross-modal alignment. Empirically, we evaluate the performance of the model with varying target modality. As shown in Table 4, language anchor outperforms visual/acoustic variants. Nevertheless, other variants of CaReFlow also obtain strong results, verifying the generalizability of CaReFlow.

6.2. Ablation Study of Batch Size

The value of batch size determines the number of modality pairs used to train the rectified flow model. In this section, we analyze the sensitivity of model performance with respect to the change of the value of batch size. As presented in Table 5, Although a larger batch size generally leads to stronger performance, the results are stable across batch sizes (±1% in Acc2), indicating CaReFlow is not sensitive to batch size because 'one-to-many' mapping ensures sufficient cross-sample pairs even with small batches.

6.3. Modality Gap Reduction

In this section, we provide a definition and quantification of the modality gap. We formalize the modality gap using Wasserstein-2 distance: $W_2(X,Y) = \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n} T_{ij}^{*} \cdot \|x_i - y_j\|_2^2}$. As shown in Table 6, CaReFlow reduces gaps more than all competitive baselines, providing direct evidence of modality gap reduction.

6.4. Comparison with Simple Transformations

We conduct additional experiments in Table 3, suggesting CaReFlow outperforms all simple transformations including contrastive learning [3], MLP, and linear layer, mainly because: (1) CaReFlow's ODE-based trajectory modeling captures smoother distribution mappings; (2) Adaptive relaxed alignment outperforms contrastive learning that treats all pairs equally; and (3) Global distribution awareness learns more effective mapping. To be more specific, unlike simple MLP/Linear layers or contrastive learning that learn point-wise mappings, CaReFlow explicitly models entire target modality distribution. The 'one-to-many' mapping allows each source data to learn from **global target distribution**, not just individual pairs, which is essential for handling scarce data. Moreover, CaReFlow learns a **straight ODE path between distributions**, enabling faster and more stable convergence.

References

[1] Tadas Baltrusaitis, 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)*, pages 59–66. IEEE, 2018. 2

[2] Santiago Castro, Devamanyu Hazarika, Verónica Pérez-Rosas, Roger Zimmermann, Rada Mihalcea, and Soujanya Poria. Towards multimodal sarcasm detection (an _obviously_ perfect paper). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4619–4629, 2019. 2

[3] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020. 5

[4] Gilles Degottex, John Kane, Thomas Drugman, Tuomo Raitio, and Stefan Scherer. Covarep: A collaborative voice analysis repository for speech technologies. In *ICASSP*, pages 960–964, 2014. 2

[5] J. Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*, 2019. 1, 2

[6] Florian Eyben. Opensmile: the munich versatile and fast open-source audio feature extractor. In *ACM International Conference on Multimedia*, pages 1459–1462, 2010. 2

[7] Yiyang Fang, Wenke Huang, Guancheng Wan, Kehua Su, and Mang Ye. Emoe: Modality-specific enhanced dynamic emotion experts. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 14314–14324, 2025. 3

[8] 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 2024*, pages 14755–14766, 2024. 2, 4

[9] Dimitris Gkoumas, Qiuchi Li, C. Lioma, Yijun Yu, and Da wei Song. What makes the difference? an empirical comparison of fusion strategies for multimodal language analysis. *Information Fusion*, 66:184–197, 2021. 2

[10] Wei Han, Hui Chen, and Soujanya Poria. Improving multimodal fusion with hierarchical mutual information maximization for multimodal sentiment analysis. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9180–9192, Online and Punta

Table 5. Batch size sensitivity analysis on CMU-MOSEI.

Batch Size	4	8	16	32	64
Acc2/MAE	87.0 / 0.515	87.5 / 0.518	87.4 / 0.509	87.6 / **0.504**	**87.9 / 0.504**

Table 6. Results of modality gap reduction on SIMS-v2 ($d = 100$).

Modality Pairs	Vanilla	MulT	ARGF	DCCA	Diffusion Bridge	CLGSI	CaReFlow
Language & Acoustic	12.70	10.86	11.61	12.19	11.52	11.99	**8.76**
Language & Visual	12.71	10.77	11.65	11.79	11.52	11.76	**8.80**
Visual & Acoustic	11.93	10.19	10.99	11.04	10.75	11.38	**8.95**

Cana, Dominican Republic, 2021. Association for Computational Linguistics. 4

[11] Md Kamrul Hasan, Wasifur Rahman, AmirAli Bagher Zadeh, Jianyuan Zhong, Md Iftekhar Tanveer, Louis-Philippe Morency, and Mohammed Ehsan Hoque. Ur-funny: A multimodal language dataset for understanding humor. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2046–2056, 2019. 1

[12] Md Kamrul Hasan, Sangwu Lee, Wasifur Rahman, Amir Zadeh, Rada Mihalcea, Louis-Philippe Morency, and Ehsan Hoque. Humor knowledge enriched transformer for understanding multimodal humor. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 12972–12980, 2021. 1, 2, 4

[13] Devamanyu Hazarika, R. Zimmermann, and Soujanya Poria. Misa: Modality-invariant and -specific representations for multimodal sentiment analysis. *ACM MM*, pages 1122–1131, 2020. 4

[14] Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. Deberta: Ddecoding-enhanced bert with disentangled attention. In *International Conference on Learning Representations*, 2021. 1, 2

[15] Changqin Huang, Jili Chen, Qionghao Huang, Shijin Wang, Yaxin Tu, and Xiaodi Huang. Atcaf: Attention-based causality-aware fusion network for multimodal sentiment analysis. *Information Fusion*, 114:102725, 2025. 3

[16] Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. Albert: A lite bert for self-supervised learning of language representations. In *International Conference on Learning Representations*, 2020. 1, 2

[17] Yong Li, Yuanzhi Wang, and Zhen Cui. Decoupled multimodal distilling for emotion recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6631–6640, 2023. 3

[18] 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: Chsims v2. 0 dataset and av-mixup consistent module. In *Proceedings of the 2022 international conference on multimodal interaction*, pages 247–258, 2022. 1, 2, 4

[19] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. 2

[20] Sijie Mai, Ya Sun, Ying Zeng, and Haifeng Hu. Excavating multimodal correlation for representation learning. *Information Fusion*, 91:542–555, 2023. 2, 4

[21] Sijie Mai, Ying Zeng, and Haifeng Hu. Multimodal information bottleneck: Learning minimal sufficient unimodal and multimodal representations. *IEEE Transactions on Multimedia*, 25:4121–4134, 2023. 3

[22] Sijie Mai, Ying Zeng, and Haifeng Hu. Learning from the global view: Supervised contrastive learning of multimodal representation. *Information Fusion*, 100:101920, 2023. 1, 2, 4

[23] Sijie Mai, Ying Zeng, Shuangjia Zheng, and Haifeng Hu. Hybrid contrastive learning of tri-modal representation for multimodal sentiment analysis. *IEEE Transactions on Affective Computing*, 14(3):2276–2289, 2023. 2

[24] Sijie Mai, Ya Sun, Aolin Xiong, Ying Zeng, and Haifeng Hu. Multimodal boosting: Addressing noisy modalities and identifying modality contribution. *IEEE Transactions on Multimedia*, 26:3018–3033, 2024. 2, 3

[25] Shraman Pramanick, Aniket Roy, and Vishal M Patel. Multimodal learning using optimal transport for sarcasm and humor detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3930–3940, 2022. 4

[26] Wasifur Rahman, M. Hasan, Sangwu Lee, Amir Zadeh, Chengfeng Mao, Louis-Philippe Morency, and E. Hoque. Integrating multimodal information in large pretrained transformers. *ACL*, 2020:2359–2369, 2020. 4

[27] Mohit Tomar, Abhisek Tiwari, Tulika Saha, and Sriparna Saha. Your tone speaks louder than your face! modality order infused multi-modal sarcasm detection. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 3926–3933, 2023. 4

[28] 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*, pages 6558–6569, 2019. 4

[29] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017. 1

[30] Pan Wang, Qiang Zhou, Yawen Wu, Tianlong Chen, and Jingtong Hu. Dlf: Disentangled-language-focused multimodal sentiment analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 21180–21188, 2025. 3

[31] Sheng Wu, Dongxiao He, Xiaobao Wang, Longbiao Wang, and Jianwu Dang. Enriching multimodal sentiment analysis through textual emotional descriptions of visual-audio content. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1601–1609, 2025. 3

[32] Xiongye Xiao, Gengshuo Liu, Gaurav Gupta, Defu Cao, Shixuan Li, Yaxing Li, Tianqing Fang, Mingxi Cheng, and Paul Bogdan. Neuro-inspired information-theoretic hierarchical perception for multimodal learning. In *The Twelfth International Conference on Learning Representations*, 2024. 1, 2, 3

[33] Zhi Xu, Dingkang Yang, Mingcheng Li, Yuzheng Wang, Zhaoyu Chen, Jiawei Chen, Jinjie Wei, and Lihua Zhang. Debiased multimodal understanding for human language sequences. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 14450–14458, 2025. 4

[34] Wenmeng Yu, Hua Xu, Ziqi Yuan, and Jiele Wu. Learning modality-specific representations with self-supervised multitask learning for multimodal sentiment analysis. In *Proceedings of the AAAI conference on artificial intelligence*, pages 10790–10797, 2021. 3

[35] Amir Zadeh, Rowan Zellers, Eli Pincus, and Louis-Philippe Morency. Multimodal sentiment intensity analysis in videos: Facial gestures and verbal messages. *IEEE Intelligent Systems*, 31(6):82–88, 2016. 1

[36] AmirAli Bagher Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2236–2246, 2018. 1

[37] Yazhou Zhang, Yang Yu, Dongming Zhao, Zuhe Li, Bo Wang, Yuexian Hou, Prayag Tiwari, and Jing Qin. Learning multitask commonness and uniqueness for multimodal sarcasm detection and sentiment analysis in conversation. *IEEE Transactions on Artificial Intelligence*, 5(3):1349–1361, 2024. 4