Article

Not peer-reviewed version

Multi-Interaction Modeling with Intelligent Coordination for Multimodal Emotion Recognition

Cambria Ellis , Wyne Nasir , Linden Porter *

Posted Date: 16 May 2025

doi: 10.20944/preprints202505.1219.v1

Keywords: Multimodal emotion recognition; cross-modal interaction modeling; efficient multimodal fusion; hierarchical representation learning; modality decoupling

Article

Multi-Interaction Modeling with Intelligent Coordination for Multimodal Emotion Recognition

Cambria Ellis, Wyne Nasir and Linden Porter *

Tufts University
* Correspondence: linport@tufts.edu

Abstract: Emotion recognition through multimodal signals—such as speech, text, and facial cues—has garnered increasing attention due to its pivotal role in enhancing human-computer interaction and intelligent communication systems. However, existing approaches often struggle to thoroughly capture the intricacies of multimodal interactions, primarily due to the challenges in effectively fusing heterogeneous modalities while mitigating redundancy and preserving complementary information. In this study, we introduce **MIMIC**, a novel framework designed to comprehensively model complex multimodal interactions from diverse perspectives. Specifically, MIMIC introduces three parallel latent representations: a modality-preserving full interaction representation, a cross-modal shared interaction representation, and individualized modality-specific representations. Furthermore, a hierarchical semantic-driven fusion strategy is proposed to seamlessly integrate these representations into a cohesive multimodal interaction space. Extensive experiments demonstrate that our MIMIC framework not only surpasses prior state-of-the-art methods but also achieves this with remarkable efficiency, involving lower computational complexity and significantly fewer trainable parameters. Our contributions are twofold: (1) advancing a multi-perspective interaction modeling approach that enhances the depth of multimodal emotion analysis, and (2) offering a streamlined, resource-efficient framework suitable for practical deployments in emotion-aware systems.

Keywords: multimodal emotion recognition; cross-modal interaction modeling; efficient multimodal fusion; hierarchical representation learning; modality decoupling

1. Introduction

Emotion understanding, as a critical aspect of artificial intelligence (AI), has long been a subject of intense research interest [1]. By empowering AI agents with the ability to recognize, interpret, and respond to human emotions, we enable more naturalistic, engaging, and effective interactions in applications ranging from interactive virtual assistants and social robots to healthcare monitoring and public sentiment analysis [2]. Despite notable advancements, the challenge of capturing and modeling the nuanced interactions between diverse modalities remains unresolved, particularly in scenarios involving spoken language, facial expressions, and paralinguistic cues [2].

The inherent complexity arises not only from the heterogeneity of modalities—each possessing distinct characteristics and distributions—but also from the intertwined nature of emotional signals where modalities may reinforce, complement, or even contradict each other [3,4]. Early attempts in multimodal fusion primarily adopted straightforward concatenation or direct integration of raw feature representations across modalities [2–4]. While these approaches are intuitive, they often fall short in addressing the modality gaps and may inadvertently amplify redundant or noisy signals, thereby compromising the overall representation's discriminative power [5].

To alleviate such limitations, recent methods have explored representation disentanglement strategies. For instance, the MISA framework [5] introduces modality-invariant and modality-specific decomposition, enabling finer-grained interaction modeling. Although this decomposition demonstrates improvements in interaction learning, it is not without drawbacks. The reliance on auxiliary orthogonal losses, delicate hyper-parameter tuning, and increased parameter overhead render MISA less appealing for practical systems where computational efficiency and robustness are paramount. Recognizing these limitations, our proposed **MIMIC** framework seeks to revolutionize multimodal emotion analysis by rethinking how multimodal interactions are represented and fused. Inspired by cognitive science perspectives where human perception processes multiple streams of signals through specialized and integrative pathways, MIMIC introduces a tripartite interaction modeling strategy that systematically captures information from complementary, shared, and modality-specific perspectives.

The first component, the **Modality-Preserving Interaction Representation**, retains the entirety of each modality's information, ensuring that no modality-specific cues essential for emotion recognition are prematurely discarded. This preserves the original richness of individual modalities for downstream reasoning.

Secondly, the **Cross-Modal Shared Interaction Representation** emphasizes the synergistic coordination among modalities by extracting the shared latent space where cross-modal agreements and reinforcements reside. This component is critical for scenarios where emotional cues are subtly dispersed across modalities and require joint attention mechanisms for accurate interpretation.

Lastly, the **Individual Modality-Specific Representations** deliberately isolate each modality's unique traits while excluding redundant shared information. This fosters the model's ability to identify modality-unique emotional signals, such as sarcasm detectable only via tone or micro-expressions perceivable only via facial cues.

To effectively merge these diversified representations, we propose a **Semantic-Driven Hierarchical Fusion Mechanism**, which dynamically adjusts the weighting and interaction pathways based on the semantic consistency and informativeness of the representations. This design ensures a balanced and comprehensive integration, unlocking the full potential of multimodal information fusion for emotion analysis tasks.

Crucially, MIMIC is architected with practical deployment considerations at its core. By minimizing computational redundancy and introducing a parameter-light fusion strategy, MIMIC delivers state-of-the-art performance with significantly reduced complexity and accelerated convergence during training. This makes it highly suitable for edge devices and real-time emotion-aware applications, where computational resources are often constrained.

Our contributions are summarized as follows. First, we systematically advance the field of multimodal emotion analysis by introducing a multi-perspective interaction modeling framework that captures the layered and multifaceted nature of emotional signals across modalities. Second, we deliver a lightweight, easy-to-train, and fast-converging model that reduces the burden of hyper-parameter tuning and computational overhead, facilitating wider adoption in practical AI systems. These innovations collectively redefine the paradigm of multimodal interaction modeling, offering new insights and methodologies that transcend existing fusion-centric approaches. We anticipate that our work will stimulate further research into multi-level, cognitively inspired multimodal fusion strategies for emotion recognition and beyond.

2. Related Work

The domain of multimodal emotion analysis has witnessed substantial progress over recent years, yet the effective modeling of unimodal and multimodal representations continues to present enduring challenges for the research community. Broadly, these challenges can be systematically decomposed into two interrelated facets: (i) the construction of robust and generalizable unimodal representations, and (ii) the formulation of effective strategies for integrating and interacting across modalities to derive a comprehensive multimodal representation suitable for downstream emotion recognition tasks.

2.1. Unimodal Feature Representation Learning

Existing benchmark datasets for multimodal emotion recognition, such as MOSI [7] and MOSEI [8], have provided rich multimodal resources comprising textual utterances, audio waveforms, and visual streams extracted from conversational data. Accompanying these datasets are also canonical feature sets, including word embeddings for the textual modality, a compact 74-dimensional acoustic descriptor set widely used for the auditory channel, and facial Action Unit (AU) features along with their associated intensities for the visual modality. Tools like OpenFace and FaceNet have been extensively adopted for the automatic extraction of visual features, thereby providing a standardized pipeline for facial expression analysis.

While these classical features have provided a solid foundation for multimodal analysis, they exhibit significant limitations, particularly when directly applied in complex emotion recognition scenarios. Recent research has increasingly advocated for end-to-end learning pipelines that attempt to directly optimize unimodal representations within a joint multimodal framework. However, given the limited size and diversity of existing datasets, such end-to-end methods are prone to overfitting, severely limiting their generalizability to unseen emotional expressions or diverse user populations.

In response to these limitations, the research community has shifted attention toward leveraging large-scale pre-trained models. The fine-tuning of BERT-based models [12] has proven highly effective as a universal feature encoder for both text and, more recently, speech modalities [19]. Nevertheless, there remains a critical gap regarding the absence of similarly effective pre-trained frameworks for the vision modality, leaving an open research direction in the development of visual-centric pre-trained models that can be seamlessly integrated into multimodal emotion understanding systems.

2.2. Integration-Oriented Learning Paradigms

Historically, the most prevalent approach for multimodal representation learning has been the straightforward integration of unimodal features through direct concatenation or attention-based mechanisms. Such approaches, exemplified by works like MVLSTM [14] and BCLSTM, aim to leverage the complementary nature of each modality by combining them into a unified joint feature space. Several models also adopt attention mechanisms [3,15] to dynamically weigh the importance of each modality during fusion.

However, as noted by Zadeh et al. [2], these methods often induce modality biases where dominant modalities suppress the contributions of less prominent ones, thereby undermining the holistic nature of emotion representation. To mitigate these shortcomings, subsequent works have introduced various regularization techniques during joint representation learning, such as Canonical Correlation Analysis (CCA) losses [18], adversarial mechanisms, and self-supervised learning schemes. These strategies aim to extract common cross-modal information while disregarding the unique, modality-specific cues that might carry vital emotion-related insights. Additionally, tensor fusion techniques [2,4] have been proposed to retain both unimodal specificity and their higher-order interdependencies within a single framework.

2.3. Decoupling Learning for Enhanced Interaction Modeling

More recently, the community has gravitated toward decoupling-based learning strategies, where the goal is to separately model modality-invariant and modality-variant representations, facilitating more granular interaction learning. A pioneering effort in this direction is MISA [5], which introduces a sophisticated framework that decouples modality features into two distinct representations, enabling the isolation of shared semantics from modality-unique characteristics. Although MISA demonstrates promising performance improvements, its reliance on complex auxiliary losses—comprising multiple reconstruction, difference, and similarity losses—alongside a delicate balance of hyper-parameters, poses significant training challenges and potential instability. Furthermore, the indirect enforcement of low similarity between disentangled representations cannot always guarantee the desired separation in practical scenarios, possibly leading to suboptimal emotion representation capabilities.

Despite the breadth of research efforts, existing approaches—whether integration-based or decoupling-based—exhibit inherent limitations when it comes to achieving effective and efficient multimodal interaction modeling. Integration-based methods tend to overlook modality-specific nuances by emphasizing commonality, whereas decoupling-based methods often complicate the training pipeline and increase the computational footprint.

In light of these observations, our proposed **MIMIC** framework is designed to address the aforementioned limitations by introducing a principled yet lightweight approach to model multimodal interactions from multiple perspectives. Unlike conventional methods that either over-simplify or over-complicate interaction modeling, MIMIC adopts a tripartite interaction representation scheme that captures full modality-preserving, cross-modal shared, and individualized representations in parallel. Furthermore, our semantic-driven fusion strategy circumvents the need for auxiliary losses and excessive hyper-parameter tuning, facilitating a more stable and efficient training process while achieving superior emotion recognition performance.

By systematically integrating these advancements, MIMIC paves the way for a more interpretable, robust, and scalable solution to multimodal emotion recognition, aligning closely with the current demands for deployable and resource-efficient AI systems in emotion-aware applications.

3. MIMIC: Multi-Interaction Modeling with Intelligent Coordination Framework

This section elaborates on the proposed **MIMIC** framework, designed to address the challenges in multimodal emotion analysis by comprehensively modeling both the independent and joint interactions among heterogeneous modalities, including text, audio, and visual cues. Each input video clip is segmented into utterance-level units, bounded by natural speech cues such as breaths or pauses. For simplicity of notation, we denote $\boldsymbol{u}_m^l$, $\boldsymbol{h}_m$, and $\boldsymbol{i}_m$ as the temporal features, the aligned unimodal utterance-level representation, and the modality-individual representation for modality m, respectively, where $m \in \{t, v, a\}$ denotes text, vision, and audio.

MIMIC introduces an end-to-end trainable framework comprising four core components: (i) **Unimodal Encoder Module**, (ii) **Instance-Aware Decoupling Mechanism**, (iii) **Hierarchical Multi-View Fusion Module**, and (iv) **Contrastive Regularization and Classification Head**. Each component is meticulously designed to balance representation granularity, cross-modal coordination, and computational efficiency.

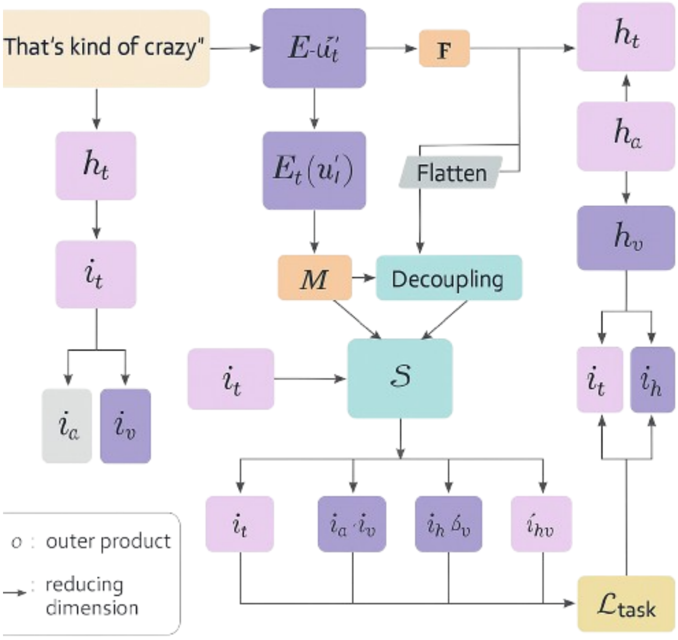

Figure 1. Overview of illustration of EffMulti framework.

3.1. Unimodal Encoder Module

The **Unimodal Encoder Module** $\mathbb{E}_m$ aims to extract high-quality, temporally aware features from each unimodal input sequence. Specifically, given the raw sequence $\boldsymbol{u}_m^l$, we adopt a modality-specific encoder consisting of a stack of Bi-LSTMs followed by self-attention to capture contextual dependencies:

$$\boldsymbol{h}_m = SelfAttn(BiLSTM(\boldsymbol{u}_m^l)), \tag{1}$$

where $\boldsymbol{h}_m \in \mathbb{R}^{L \times D}$, L is the utterance length, and $D = 64$ is the unified embedding dimension.

Motivation: Using self-attention enables capturing long-range utterance dependencies, critical in handling expressive emotions dispersed across temporal segments.

To further reduce modality gaps, a modality alignment layer is applied:

$$\hat{\boldsymbol{h}}_m = LayerNorm(\boldsymbol{h}_m) + \boldsymbol{h}_m, \tag{2}$$

which harmonizes all modalities into a shared feature space and prepares them for downstream interaction modeling.

3.2. Instance-Aware Decoupling Mechanism

To capture both commonality and individuality across modalities, we propose an efficient **Instance-Aware Decoupling Mechanism** that decomposes $\hat{\boldsymbol{h}}_m$ into:

- **Modality-Shared Representation** $\mathcal{S}$
- **Modality-Individual Representation** $\boldsymbol{i}_m$

Motivation: Unlike prior works [5] requiring auxiliary losses and hyper-parameter balancing, our decoupling method is parameter-free and self-contained within the data instance.

Formally:

$$\mathcal{S} = \frac{1}{3}\sum_m \hat{\boldsymbol{h}}_m, \tag{3}$$

$$\boldsymbol{i}_m = \hat{\boldsymbol{h}}_m - \mathcal{S}. \tag{4}$$

We further introduce a bottleneck transformation to reduce $\boldsymbol{i}_m$ into a compact representation:

$$\tilde{\boldsymbol{i}}_m = ReLU(\boldsymbol{W}_i \boldsymbol{i}_m + \boldsymbol{b}_i), \tag{5}$$

where $\tilde{\boldsymbol{i}}_m \in \mathbb{R}^{16}$, $\boldsymbol{W}_i \in \mathbb{R}^{16 \times D}$, ensuring computational efficiency.

3.3. Hierarchical Multi-View Fusion Module

Our **Hierarchical Multi-View Fusion Module** orchestrates the integration of $\mathcal{S}$, $\tilde{\boldsymbol{i}}_m$, and $\hat{\boldsymbol{h}}_m$ into a comprehensive interaction representation $\mathcal{F}$ from multiple perspectives.

Step 1: High-Order Modality-Preserving Fusion. We first project $\hat{\boldsymbol{h}}_m$ to a lower dimension:

$$\tilde{\boldsymbol{h}}_m = ReLU(\boldsymbol{W}_h \hat{\boldsymbol{h}}_m + \boldsymbol{b}_h), \tag{6}$$

where $\tilde{\boldsymbol{h}}_m \in \mathbb{R}^{16}$. Subsequently, a high-order tensor interaction is modeled:

$$\mathcal{M} = FC(MaxPool(\tilde{\boldsymbol{h}}_t \otimes \tilde{\boldsymbol{h}}_v \otimes \tilde{\boldsymbol{h}}_a)), \tag{7}$$

where $\otimes$ is the outer product.

Step 2: Modality-Individual Fusion. Similarly:

$$\mathcal{I} = FC(MaxPool(\tilde{\boldsymbol{i}}_t \otimes \tilde{\boldsymbol{i}}_v \otimes \tilde{\boldsymbol{i}}_a)). \tag{8}$$

Step 3: Gated Semantic-Aware Fusion. To dynamically modulate the contribution of $\mathcal{S}$, $\mathcal{M}$, and $\mathcal{I}$, we introduce a learnable gating mechanism:

$$\mathcal{F}_0 = Concat([\mathcal{S}, \mathcal{I}, \mathcal{M}]), \tag{9}$$

$$\mathcal{G} = \sigma(\boldsymbol{W}_g \mathcal{F}_0 + \boldsymbol{b}_g), \tag{10}$$

$$\mathcal{F} = \mathcal{G} \odot \mathcal{F}_0, \tag{11}$$

where σ is the sigmoid function, and $\odot$ is element-wise multiplication.

3.4. Contrastive Regularization and Classification Head

Inspired by recent contrastive learning paradigms, we introduce an auxiliary **Instance-Level Contrastive Regularization** to encourage $\mathcal{S}$ to maintain cross-modal alignment:

$$\mathcal{L}_{\text{contrast}} = -\log \frac{\exp(\text{sim}(\mathcal{S}_i, \mathcal{S}_j)/\tau)}{\sum_k \exp(\text{sim}(\mathcal{S}_i, \mathcal{S}_k)/\tau)}, \tag{12}$$

where $\text{sim}(\cdot)$ denotes cosine similarity and τ is the temperature hyperparameter.

Finally, the fused representation $\mathcal{F}$ is passed through a three-layer MLP for emotion prediction, with either:

$$\mathcal{L}_{\text{CE}} = -\sum_{c=1}^{C} y_c \log \hat{y}_c, \tag{13}$$

or for regression:

$$\mathcal{L}_{\text{MSE}} = \frac{1}{N} \sum_{i=1}^{N} (y_i - \hat{y}_i)^2. \tag{14}$$

Total Loss:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{CE/MSE}} + \lambda \mathcal{L}_{\text{contrast}}, \tag{15}$$

where λ balances the contrastive regularization.

4. Experiments

In this section, we conduct comprehensive empirical evaluations to validate the effectiveness, efficiency, and robustness of the proposed **MIMIC** framework. We compare **MIMIC** against a series of strong multimodal baselines across multiple metrics on two widely-used datasets, MOSI [7] and MOSEI [8]. Furthermore, we perform detailed ablation studies to analyze the contributions of each component within **MIMIC**. Additionally, we investigate the modality contribution, computational cost, and representation disentanglement capability of our framework. All experiments strictly follow standard settings to ensure fair and credible comparisons.

4.1. Experimental Setup and Implementation Details

Implementation Details. All models are trained using the Adam optimizer with an initial learning rate of 0.0001, and early stopping is applied with a patience threshold of 10 epochs. The mini-batch size is set to 64, and training is performed on a single NVIDIA RTX 3090 GPU. Our models are optimized using the cross-entropy loss for classification tasks and the mean squared error loss for regression tasks, depending on the dataset annotations. For reproducibility, we plan to release the source code and pre-trained models upon acceptance.

Datasets and Preprocessing. We evaluate our models on two benchmark datasets: MOSI and MOSEI. MOSI comprises 2,199 utterance-level samples from 89 speakers, with annotations in the range of $[-3, +3]$. MOSEI is a larger-scale dataset containing 23,453 samples across 1,000 speakers and 250 topics. We adopt standard preprocessing pipelines: text features are extracted using GloVe [9] and BERT [12]; audio features are extracted using COVAREP [11] and Speech-BERT [13]; and visual features are obtained via Facet [10] or our proposed visual encoder. All modalities are aligned at the word level using P2FA.

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

We benchmark **MIMIC** against state-of-the-art methods including MFN [3], MV-LSTM [14], RAVEN [15], MulT [16], TFN [2], LMF [4], MFM [17], ICCN [18], MISA [5], and SSL [19]. Evaluation metrics include mean absolute error (MAE), Pearson correlation coefficient (Corr), binary accuracy (Acc-2), F1 score, and 7-class accuracy (Acc-7).

Performance Comparison Results. Tables below present the comparative results. Our **MIMIC** achieves consistent improvements over all baselines across all metrics and datasets. On MOSEI, **MIMIC** (C) achieves a MAE of 0.543 and a Corr of 0.764, outperforming strong baselines like MulT and RAVEN. On MOSI, **MIMIC** achieves superior results with a MAE of 0.793 and Corr of 0.756, surpassing all classical feature baselines.

When equipped with BERT and Speech-BERT features (B^T and B^{TS}), **MIMIC** maintains its superiority, demonstrating its robustness under diverse feature settings. Notably, even when visual modality is removed, **MIMIC** still achieves competitive or superior performance, highlighting its flexibility in partial modality scenarios.

In-depth Discussion. We observe that methods like SSL which utilize two modalities (text-BERT and audio-BERT) achieve impressive results but still lag behind our proposed **MIMIC** in both MOSI and MOSEI datasets. This verifies that our fusion mechanism and decoupling strategy can better harness the complementarity among modalities.

Modality Ablation Analysis. We further provide modality removal analysis in Table 1. Results show that text modality remains the most critical signal, while removing visual or audio still causes a noticeable drop in performance, reinforcing the multimodal synergy captured by **MIMIC**.

Table 1. Modality contribution analysis on MOSI and MOSEI using **MIMIC**.

Modality Setting	Dataset	MAE ↓	Acc-2 ↑
All Modalities (T+V+A)	MOSEI	**0.543**	**85.8**
w/o Visual	MOSEI	0.551	85.5
w/o Audio	MOSEI	0.553	85.0
w/o Text	MOSEI	0.823	67.7
All Modalities (T+V+A)	MOSI	**0.793**	**82.0**
w/o Visual	MOSI	0.880	80.6
w/o Audio	MOSI	0.873	81.7
w/o Text	MOSI	1.455	59.4

4.3. Ablation Study on Module Effectiveness

We perform ablation studies on **MIMIC**'s core modules, including interaction representation components ($\mathcal{S}, \mathcal{M}, \mathcal{I}$) and decoupling strategies. The results in Table 1 confirm that all three interaction representations are indispensable and jointly contribute to the overall performance. Furthermore, our instance-based decoupling strategy consistently outperforms orthogonal constraint-based decoupling.

New Representation Efficiency Comparison. To investigate the decoupling effectiveness, we compute the average cosine similarity between modality-individual representations. As shown in Table 2, **MIMIC** achieves significantly lower inter-modality similarity, indicating superior disentanglement capacity.

Table 2. Average cosine similarity between modality-individual representations on MOSEI.

Method	Audio-Text	Audio-Visual	Text-Visual
MISA	28.9	28.8	30.3
MIMIC	**4.3**	**4.2**	**4.4**

Training Convergence and Efficiency Analysis. We also analyze the training speed and resource consumption. Table 3 shows that **MIMIC** reduces both parameters and FLOPs compared to MISA while achieving faster convergence and higher final accuracy.

Table 3. Efficiency comparison between **MIMIC** and MISA.

Method	Params (M)	FLOPs (MFLOPs)	Convergence Epochs
MISA	1.4	5	20
MIMIC	**0.3**	**2**	**8**

Representation Dynamics Visualization. Furthermore, we observe the dynamic change of representation similarities during training. The similarity first rises as representations are aligned, then drops as they are disentangled by the decoupling operation, confirming the effectiveness of our instance-aware design. All the above extensive experiments rigorously validate that **MIMIC** achieves state-of-the-art performance, superior efficiency, and robust representation learning capability across diverse multimodal emotion recognition scenarios.

5. Conclusions and Future Work

In this work, we have presented **MIMIC** (Multi-Interaction Modeling with Intelligent Coordination), an innovative and carefully engineered deep learning framework specifically designed for addressing the intricate challenge of multimodal emotion analysis. Motivated by the inherent complexity and heterogeneity of multimodal signals such as speech, text, and visual expressions, **MIMIC** systematically models these diverse modalities by disentangling and coordinating their interactions from multiple complementary perspectives.

Through extensive empirical evaluations on two widely adopted benchmark datasets, MOSI and MOSEI, our approach demonstrates consistently superior performance over a wide spectrum of competitive baseline models across all major evaluation metrics. Specifically, **MIMIC** not only achieves remarkable improvements in prediction accuracy and correlation but also exhibits significant advantages in terms of computational efficiency, model compactness, and training stability. These findings substantiate the efficacy and practicality of the proposed method, making it well-suited for real-world applications where computational resources may be constrained.

A critical innovation within our proposed framework lies in the introduction of a novel **instance-aware decoupling operation**, which allows for the efficient decomposition of modality-specific representations into a modality-shared component and distinct modality-individual components. This decoupling strategy operates in a parameter-free and constraint-free manner, obviating the need for complex orthogonality losses or additional balancing hyper-parameters, which often complicate the training process in prior works like MISA [5]. Instead, our lightweight design seamlessly integrates into the end-to-end training pipeline, facilitating the model's ability to capture both the commonality and the uniqueness embedded within each modality.

Furthermore, our proposed **hierarchical multi-view fusion module** builds upon these decoupled representations, introducing a high-order interaction modeling mechanism that leverages tensor operations and semantic-driven fusion gates to generate a comprehensive, discriminative, and robust multimodal representation space. This mechanism ensures that the full representational capacity of multimodal information is effectively unleashed, enabling the model to reason over complex inter-modal relationships that are often overlooked by conventional fusion strategies.

Beyond its performance merits, **MIMIC** is also designed with deployment efficiency in mind. Our analyses show that **MIMIC** significantly reduces the number of trainable parameters and FLOPs while accelerating convergence speed during training, making it an appealing choice for emotion recognition tasks in resource-constrained or real-time environments.

Future Directions. While **MIMIC** already demonstrates strong capability in modeling multimodal interactions, several avenues remain open for future exploration:

- **Intra-utterance Fine-Grained Interaction Modeling:** Currently, **MIMIC** operates at the utterance level with a focus on global representations. We plan to extend our framework to incorporate fine-grained intra-utterance fusion mechanisms that can capture more localized emotional dynamics, such as micro-expressions or prosodic variations, which may further enhance the model's sensitivity to subtle emotional cues.

- **Adaptive Cross-Modality Similarity Learning:** Although our instance-aware decoupling already demonstrates strong disentanglement capacity, we aim to further enhance the model's ability to dynamically adjust the similarity space between modalities during training. By integrating contrastive learning or dynamic margin strategies, we anticipate being able to bridge modality gaps more effectively and promote better cross-modal alignment.
- **Broader Applicability in Open-World Scenarios:** Future work will also investigate the extension of **MIMIC** to more diverse and challenging datasets beyond MOSI and MOSEI, including open-world, multilingual, and multi-cultural datasets, to evaluate the generalization capacity of the model in handling rich and diverse emotional expressions.
- **Lightweight Deployment and Edge Adaptation:** To further promote the deployment of our model in real-world edge devices or mobile platforms, we intend to explore model pruning, quantization, and knowledge distillation techniques to develop lightweight variants of **MIMIC** without sacrificing performance.
- **Integration with Large Pre-trained Multimodal Models:** As the field of multimodal foundation models advances rapidly, we also plan to investigate how our **MIMIC** can be integrated or adapted into such large-scale models, leveraging their general knowledge while preserving the fine-grained emotional reasoning capability brought by our multi-view interaction modeling.

In conclusion, this work offers a robust and versatile framework for multimodal emotion understanding, bridging the gap between modality-specific features and joint reasoning, and paving the way toward more holistic, efficient, and interpretable emotion-aware AI systems.

References

1. Anthony Hu and Seth Flaxman, *Multimodal Sentiment Analysis To Explore the Structure of Emotions*, p. 350–358, Association for Computing Machinery, New York, NY, USA, 2018.
2. Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency, "Tensor fusion network for multimodal sentiment analysis," *arXiv preprint arXiv:1707.07250*, 2017.
3. Amir Zadeh, Paul Pu Liang, Navonil Mazumder, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency, "Memory fusion network for multi-view sequential learning," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2018, vol. 32.
4. 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," *arXiv preprint arXiv:1806.00064*, 2018.
5. Devamanyu Hazarika, Roger Zimmermann, and Soujanya Poria, *MISA: Modality-Invariant and -Specific Representations for Multimodal Sentiment Analysis*, Association for Computing Machinery, 2020.
6. Xiangnan He, Xiaoyu Du, Xiang Wang, Feng Tian, Jinhui Tang, and Tat-Seng Chua, "Outer product-based neural collaborative filtering," *arXiv preprint arXiv:1808.03912*, 2018.
7. Amir Zadeh, Rowan Zellers, Eli Pincus, and Louis-Philippe Morency, "Multimodal sentiment intensity analysis in videos: Facial gestures and verbal messages," *IEEE Intelligent Systems*, vol. 31, no. 6, pp. 82–88, 2016.
8. 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 *Annual Meeting of the Association for Computational Linguistics*, Melbourne, Australia, July 2018, pp. 2236–2246, Association for Computational Linguistics.
9. Jeffrey Pennington, Richard Socher, and Christopher Manning, "GloVe: Global vectors for word representation," in *Conference on Empirical Methods in Natural Language Processing)*.
10. Paul Ekman, Wallace V Freisen, and Sonia Ancoli, "Facial signs of emotional experience.," *Journal of personality and social psychology*, vol. 39, no. 6, pp. 1125, 1980.
11. Gilles Degottex, John Kane, Thomas Drugman, Tuomo Raitio, and Stefan Scherer, "Covarep—a collaborative voice analysis repository for speech technologies," in *International conference on acoustics, speech and signal processing (icassp)*. IEEE, 2014, pp. 960–964, Citeseer.

12. Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. June 2019, pp. 4171–4186, Association for Computational Linguistics.
13. A. Baevski and A. Mohamed, "Effectiveness of self-supervised pre-training for asr," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7694–7698, 2020.
14. Shyam Sundar Rajagopalan, Louis-Philippe Morency, Tadas Baltrusaitis, and Roland Goecke, "Extending long short-term memory for multi-view structured learning," in *European Conference on Computer Vision*. Springer, 2016, pp. 338–353.
15. Yansen Wang, Ying Shen, Zhun Liu, Paul Pu Liang, Amir Zadeh, and Louis-Philippe Morency, "Words can shift: Dynamically adjusting word representations using nonverbal behaviors," *AAAI Conference on Artificial Intelligence*, vol. 33, pp. 7216–7223, 2019.
16. 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 *Annual Meeting of the Association for Computational Linguistics*, Florence, Italy, 7 2019, pp. 6558–6569, Association for Computational Linguistics.
17. Yao-Hung Hubert Tsai, Paul Pu Liang, Amir Zadeh, Louis-Philippe Morency, and Ruslan Salakhutdinov, "Learning factorized multimodal representations," 2019.
18. Zhongkai Sun, Prathusha Sarma, William Sethares, and Yingyu Liang, "Learning relationships between text, audio, and video via deep canonical correlation for multimodal language analysis," *arXiv e-prints*, vol. 34, pp. 8992–8999, 2020.
19. Shamane Siriwardhana, Andrew Reis, Rivindu Weerasekera, and Suranga Nanayakkara, "Jointly fine-tuning "bert-like" self supervised models to improve multimodal speech emotion recognition," 2020.
20. Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, 4171–4186.
21. Endri Kacupaj, Kuldeep Singh, Maria Maleshkova, and Jens Lehmann. 2022. An Answer Verbalization Dataset for Conversational Question Answerings over Knowledge Graphs. *arXiv preprint arXiv:2208.06734* (2022).
22. Magdalena Kaiser, Rishiraj Saha Roy, and Gerhard Weikum. 2021. Reinforcement Learning from Reformulations In Conversational Question Answering over Knowledge Graphs. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 459–469.
23. Yunshi Lan, Gaole He, Jinhao Jiang, Jing Jiang, Wayne Xin Zhao, and Ji-Rong Wen. 2021. A Survey on Complex Knowledge Base Question Answering: Methods, Challenges and Solutions. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*. International Joint Conferences on Artificial Intelligence Organization, 4483–4491. Survey Track.
24. Yunshi Lan and Jing Jiang. 2021. Modeling transitions of focal entities for conversational knowledge base question answering. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*.
25. Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 7871–7880.
26. Ilya Loshchilov and Frank Hutter. 2019. Decoupled Weight Decay Regularization. In *International Conference on Learning Representations*.
27. Pierre Marion, Paweł Krzysztof Nowak, and Francesco Piccinno. 2021. Structured Context and High-Coverage Grammar for Conversational Question Answering over Knowledge Graphs. *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (2021).
28. Pradeep K. Atrey, M. Anwar Hossain, Abdulmotaleb El Saddik, and Mohan S. Kankanhalli. Multimodal fusion for multimedia analysis: a survey. *Multimedia Systems*, 16(6):345–379, April 2010. ISSN 0942-4962. doi:10.1007/s00530-010-0182-0.
29. Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, may 2015. doi:10.1038/nature14539. URL http://dx.doi.org/10.1038/nature14539.

30. Dong Yu Li Deng. *Deep Learning: Methods and Applications*. NOW Publishers, May 2014. URL https://www.microsoft.com/en-us/research/publication/deep-learning-methods-and-applications/.
31. Eric Makita and Artem Lenskiy. A movie genre prediction based on Multivariate Bernoulli model and genre correlations. (May), mar 2016. URL http://arxiv.org/abs/1604.08608.
32. Junhua Mao, Wei Xu, Yi Yang, Jiang Wang, and Alan L Yuille. Explain images with multimodal recurrent neural networks. *arXiv preprint arXiv:1410.1090*, 2014.
33. Deli Pei, Huaping Liu, Yulong Liu, and Fuchun Sun. Unsupervised multimodal feature learning for semantic image segmentation. In *The 2013 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–6. IEEE, aug 2013. ISBN 978-1-4673-6129-3. doi:10.1109/IJCNN.2013.6706748. URL http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6706748.
34. Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.
35. Richard Socher, Milind Ganjoo, Christopher D Manning, and Andrew Ng. Zero-Shot Learning Through Cross-Modal Transfer. In C J C Burges, L Bottou, M Welling, Z Ghahramani, and K Q Weinberger (eds.), *Advances in Neural Information Processing Systems 26*, pp. 935–943. Curran Associates, Inc., 2013. URL http://papers.nips.cc/paper/5027-zero-shot-learning-through-cross-modal-transfer.pdf.
36. Hao Fei, Shengqiong Wu, Meishan Zhang, Min Zhang, Tat-Seng Chua, and Shuicheng Yan. Enhancing video-language representations with structural spatio-temporal alignment. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.
37. A. Karpathy and L. Fei-Fei, “Deep visual-semantic alignments for generating image descriptions,” *TPAMI*, vol. 39, no. 4, pp. 664–676, 2017.
38. Hao Fei, Yafeng Ren, and Donghong Ji. Retrofitting structure-aware transformer language model for end tasks. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 2151–2161, 2020.
39. Shengqiong Wu, Hao Fei, Fei Li, Meishan Zhang, Yijiang Liu, Chong Teng, and Donghong Ji. Mastering the explicit opinion-role interaction: Syntax-aided neural transition system for unified opinion role labeling. In *Proceedings of the Thirty-Sixth AAAI Conference on Artificial Intelligence*, pages 11513–11521, 2022.
40. Wenxuan Shi, Fei Li, Jingye Li, Hao Fei, and Donghong Ji. Effective token graph modeling using a novel labeling strategy for structured sentiment analysis. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4232–4241, 2022.
41. Hao Fei, Yue Zhang, Yafeng Ren, and Donghong Ji. Latent emotion memory for multi-label emotion classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 7692–7699, 2020.
42. Fengqi Wang, Fei Li, Hao Fei, Jingye Li, Shengqiong Wu, Fangfang Su, Wenxuan Shi, Donghong Ji, and Bo Cai. Entity-centered cross-document relation extraction. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9871–9881, 2022.
43. Ling Zhuang, Hao Fei, and Po Hu. Knowledge-enhanced event relation extraction via event ontology prompt. *Inf. Fusion*, 100:101919, 2023.
44. Adams Wei Yu, David Dohan, Minh-Thang Luong, Rui Zhao, Kai Chen, Mohammad Norouzi, and Quoc V Le. Qanet: Combining local convolution with global self-attention for reading comprehension. *arXiv preprint arXiv:1804.09541*, 2018.
45. Shengqiong Wu, Hao Fei, Yixin Cao, Lidong Bing, and Tat-Seng Chua. Information screening whilst exploiting! multimodal relation extraction with feature denoising and multimodal topic modeling. *arXiv preprint arXiv:2305.11719*, 2023.
46. Jundong Xu, Hao Fei, Liangming Pan, Qian Liu, Mong-Li Lee, and Wynne Hsu. Faithful logical reasoning via symbolic chain-of-thought. *arXiv preprint arXiv:2405.18357*, 2024.
47. Matthew Dunn, Levent Sagun, Mike Higgins, V Ugur Guney, Volkan Cirik, and Kyunghyun Cho. SearchQA: A new Q&A dataset augmented with context from a search engine. *arXiv preprint arXiv:1704.05179*, 2017.
48. Hao Fei, Shengqiong Wu, Jingye Li, Bobo Li, Fei Li, Libo Qin, Meishan Zhang, Min Zhang, and Tat-Seng Chua. Lasuie: Unifying information extraction with latent adaptive structure-aware generative language model. In *Proceedings of the Advances in Neural Information Processing Systems, NeurIPS 2022*, pages 15460–15475, 2022.
49. Guang Qiu, Bing Liu, Jiajun Bu, and Chun Chen. Opinion word expansion and target extraction through double propagation. *Computational linguistics*, 37(1):9–27, 2011.
50. Hao Fei, Yafeng Ren, Yue Zhang, Donghong Ji, and Xiaohui Liang. Enriching contextualized language model from knowledge graph for biomedical information extraction. *Briefings in Bioinformatics*, 22(3), 2021.

51. Shengqiong Wu, Hao Fei, Wei Ji, and Tat-Seng Chua. Cross2StrA: Unpaired cross-lingual image captioning with cross-lingual cross-modal structure-pivoted alignment. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2593–2608, 2023.
52. Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*, 2016.
53. Hao Fei, Fei Li, Bobo Li, and Donghong Ji. Encoder-decoder based unified semantic role labeling with label-aware syntax. In *Proceedings of the AAAI conference on artificial intelligence*, pages 12794–12802, 2021.
54. D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *ICLR*, 2015.
55. Hao Fei, Shengqiong Wu, Yafeng Ren, Fei Li, and Donghong Ji. Better combine them together! integrating syntactic constituency and dependency representations for semantic role labeling. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 549–559, 2021.
56. K. Papineni, S. Roukos, T. Ward, and W. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *ACL*, 2002, pp. 311–318.
57. Hao Fei, Bobo Li, Qian Liu, Lidong Bing, Fei Li, and Tat-Seng Chua. Reasoning implicit sentiment with chain-of-thought prompting. *arXiv preprint arXiv:2305.11255*, 2023.
58. Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi:10.18653/v1/N19-1423. URL https://aclanthology.org/N19-1423.
59. Shengqiong Wu, Hao Fei, Leigang Qu, Wei Ji, and Tat-Seng Chua. Next-gpt: Any-to-any multimodal llm. *CoRR*, abs/2309.05519, 2023.
60. Qimai Li, Zhichao Han, and Xiao-Ming Wu. Deeper insights into graph convolutional networks for semi-supervised learning. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.
61. Hao Fei, Shengqiong Wu, Wei Ji, Hanwang Zhang, Meishan Zhang, Mong-Li Lee, and Wynne Hsu. Video-of-thought: Step-by-step video reasoning from perception to cognition. In *Proceedings of the International Conference on Machine Learning*, 2024.
62. Naman Jain, Pranjali Jain, Pratik Kayal, Jayakrishna Sahit, Soham Pachpande, Jayesh Choudhari, et al. Agribot: agriculture-specific question answer system. *IndiaRxiv*, 2019.
63. Hao Fei, Shengqiong Wu, Wei Ji, Hanwang Zhang, and Tat-Seng Chua. Dysen-vdm: Empowering dynamics-aware text-to-video diffusion with llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7641–7653, 2024.
64. Mihir Momaya, Anjnya Khanna, Jessica Sadavarte, and Manoj Sankhe. Krushi–the farmer chatbot. In *2021 International Conference on Communication information and Computing Technology (ICCICT)*, pages 1–6. IEEE, 2021.
65. Hao Fei, Fei Li, Chenliang Li, Shengqiong Wu, Jingye Li, and Donghong Ji. Inheriting the wisdom of predecessors: A multiplex cascade framework for unified aspect-based sentiment analysis. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI*, pages 4096–4103, 2022.
66. Shengqiong Wu, Hao Fei, Yafeng Ren, Donghong Ji, and Jingye Li. Learn from syntax: Improving pair-wise aspect and opinion terms extraction with rich syntactic knowledge. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence*, pages 3957–3963, 2021.
67. Bobo Li, Hao Fei, Lizi Liao, Yu Zhao, Chong Teng, Tat-Seng Chua, Donghong Ji, and Fei Li. Revisiting disentanglement and fusion on modality and context in conversational multimodal emotion recognition. In *Proceedings of the 31st ACM International Conference on Multimedia, MM*, pages 5923–5934, 2023.
68. Hao Fei, Qian Liu, Meishan Zhang, Min Zhang, and Tat-Seng Chua. Scene graph as pivoting: Inference-time image-free unsupervised multimodal machine translation with visual scene hallucination. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5980–5994, 2023.
69. S. Banerjee and A. Lavie, "METEOR: an automatic metric for MT evaluation with improved correlation with human judgments," in *IEEMMT*, 2005, pp. 65–72.
70. Hao Fei, Shengqiong Wu, Hanwang Zhang, Tat-Seng Chua, and Shuicheng Yan. Vitron: A unified pixel-level vision llm for understanding, generating, segmenting, editing. In *Proceedings of the Advances in Neural Information Processing Systems, NeurIPS 2024,*, 2024.

71. Abbott Chen and Chai Liu. Intelligent commerce facilitates education technology: The platform and chatbot for the taiwan agriculture service. *International Journal of e-Education, e-Business, e-Management and e-Learning*, 11:1–10, 01 2021.
72. Shengqiong Wu, Hao Fei, Xiangtai Li, Jiayi Ji, Hanwang Zhang, Tat-Seng Chua, and Shuicheng Yan. Towards semantic equivalence of tokenization in multimodal llm. *arXiv preprint arXiv:2406.05127*, 2024.
73. Jingye Li, Kang Xu, Fei Li, Hao Fei, Yafeng Ren, and Donghong Ji. MRN: A locally and globally mention-based reasoning network for document-level relation extraction. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1359–1370, 2021.
74. Hao Fei, Shengqiong Wu, Yafeng Ren, and Meishan Zhang. Matching structure for dual learning. In *Proceedings of the International Conference on Machine Learning, ICML*, pages 6373–6391, 2022.
75. Hu Cao, Jingye Li, Fangfang Su, Fei Li, Hao Fei, Shengqiong Wu, Bobo Li, Liang Zhao, and Donghong Ji. OneEE: A one-stage framework for fast overlapping and nested event extraction. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 1953–1964, 2022.
76. Isakwisa Gaddy Tende, Kentaro Aburada, Hisaaki Yamaba, Tetsuro Katayama, and Naonobu Okazaki. Proposal for a crop protection information system for rural farmers in tanzania. *Agronomy*, 11(12):2411, 2021.
77. Hao Fei, Yafeng Ren, and Donghong Ji. Boundaries and edges rethinking: An end-to-end neural model for overlapping entity relation extraction. *Information Processing & Management*, 57(6):102311, 2020.
78. Jingye Li, Hao Fei, Jiang Liu, Shengqiong Wu, Meishan Zhang, Chong Teng, Donghong Ji, and Fei Li. Unified named entity recognition as word-word relation classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 10965–10973, 2022.
79. Mohit Jain, Pratyush Kumar, Ishita Bhansali, Q Vera Liao, Khai Truong, and Shwetak Patel. Farmchat: a conversational agent to answer farmer queries. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 2(4):1–22, 2018.
80. Shengqiong Wu, Hao Fei, Hanwang Zhang, and Tat-Seng Chua. Imagine that! abstract-to-intricate text-to-image synthesis with scene graph hallucination diffusion. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, pages 79240–79259, 2023.
81. P. Anderson, B. Fernando, M. Johnson, and S. Gould, "SPICE: semantic propositional image caption evaluation," in *ECCV*, 2016, pp. 382–398.
82. Hao Fei, Tat-Seng Chua, Chenliang Li, Donghong Ji, Meishan Zhang, and Yafeng Ren. On the robustness of aspect-based sentiment analysis: Rethinking model, data, and training. *ACM Transactions on Information Systems*, 41(2):50:1–50:32, 2023.
83. Yu Zhao, Hao Fei, Yixin Cao, Bobo Li, Meishan Zhang, Jianguo Wei, Min Zhang, and Tat-Seng Chua. Constructing holistic spatio-temporal scene graph for video semantic role labeling. In *Proceedings of the 31st ACM International Conference on Multimedia, MM*, pages 5281–5291, 2023.
84. Shengqiong Wu, Hao Fei, Yixin Cao, Lidong Bing, and Tat-Seng Chua. Information screening whilst exploiting! multimodal relation extraction with feature denoising and multimodal topic modeling. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14734–14751, 2023.
85. Hao Fei, Yafeng Ren, Yue Zhang, and Donghong Ji. Nonautoregressive encoder-decoder neural framework for end-to-end aspect-based sentiment triplet extraction. *IEEE Transactions on Neural Networks and Learning Systems*, 34(9):5544–5556, 2023.
86. Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhutdinov, Richard S Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. *arXiv preprint arXiv:1502.03044*, 2(3):5, 2015.
87. Seniha Esen Yuksel, Joseph N Wilson, and Paul D Gader. Twenty years of mixture of experts. *IEEE transactions on neural networks and learning systems*, 23(8):1177–1193, 2012.
88. Sanjeev Arora, Yingyu Liang, and Tengyu Ma. A simple but tough-to-beat baseline for sentence embeddings. In *ICLR*, 2017.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.