Article

Not peer-reviewed version

Harmonious Multi-Grained Integration for Robust Multimodal Emotion Analysis

Dorian Sinclair , Emily Marwood , Caspian George * , Elodie Beaumont

Posted Date: 21 February 2025

doi: 10.20944/preprints202502.1712.v1

Keywords: multimodal emotion analysis; multi-granularity fusion; transformer architecture; cross-modal integration

Article

Harmonious Multi-Grained Integration for Robust Multimodal Emotion Analysis

Dorian Sinclair, Emily Marwood, Caspian George * and Elodie Beaumont

Flinders University
* Correspondence: caspian_george@flinders.edu.au

Abstract: In this paper, we introduce **HarmoFusion**, a comprehensive framework that seamlessly integrates multi-granular information for robust multimodal emotion analysis. Traditional approaches in emotion recognition often rely solely on either holistic, pre-trained utterance-level embeddings or isolated fine-grained features, which can result in suboptimal performance when dealing with the subtle and dynamic nature of emotional expression. HarmoFusion bridges this gap by unifying pre-trained global representations with detailed interactions at the phoneme and word levels. Inspired by advancements in transformer-based text-to-speech systems, our model employs a hierarchical attention mechanism to capture intricate cross-modal dependencies. More specifically, our architecture fuses phonetic details and lexical semantics via a novel transformer module that computes interactions. Additionally, we introduce further formulations to model the integration of word-level and phoneme-level embeddings. Extensive experiments conducted on the IEMOCAP dataset demonstrate that HarmoFusion not only surpasses current state-of-the-art methods in terms of accuracy and robustness but also exhibits enhanced performance when incorporating fine-grained interactions. Our ablation studies further highlight the importance of each component in capturing the complex nuances of multimodal emotional signals, ultimately paving the way for more effective human-computer interaction systems.

Keywords: multimodal emotion analysis; multi-granularity fusion; transformer architecture; cross-modal integration

1. Introduction

Speech Emotion Recognition (SER) has long been a focal point for developing advanced human-computer interaction systems, where the primary objective is to accurately infer a speaker's emotional state—be it happiness, anger, sadness, or any other affective condition—from their speech signals [1]. Despite considerable progress, SER faces two primary challenges. First, the acquisition of large-scale, high-quality labeled data is severely hampered by the inherent subjectivity of emotion annotation; multi-person evaluation is often required, which not only is time-consuming but also introduces inter-annotator variability [2]. Second, emotional expressions are naturally multimodal and intricately fine-grained, involving subtle cues dispersed across different levels of speech and language [3]. This necessitates a fusion of heterogeneous data sources that can capture both global and local information.

A prevalent strategy to mitigate the scarcity of annotated data is the adoption of transfer learning. Self-supervised learning (SSL) techniques have recently set new performance records in domains such as natural language processing (NLP) [4–6] and speech recognition [7–9] by pre-training models on massive amounts of unlabeled data. In the realm of emotion recognition, works by F. A. Acheampong et al. [10] and L. Pepino et al. [11] have successfully leveraged pre-trained models for text and speech modalities, respectively. Nonetheless, these methodologies typically focus on a single modality, thereby overlooking the synergistic potential of combining diverse sources of emotional cues.

To date, a wide range of methods have been explored to learn the interaction between different modalities. For instance, early fusion and late fusion techniques have been studied extensively using pre-trained models such as BERT [4] and Wav2vec [7] by researchers like S. Siriwardhana et al. [12] and Z. Zhao et al. [2]. Although late fusion strategies often demonstrate superior performance, they predominantly rely on holistic, aggregated embeddings that may not capture the essential fine-grained details required for precise emotion detection. In contrast, earlier approaches that eschewed pre-trained embeddings—such as those developed by S. Yoon et al. [13], H. Xu et al. [14], and H. Li et al. [3]—have attempted to fuse vocal and textual features using recurrent architectures and alignment techniques. However, these methods often rely on utterance-level fusion or require complex alignment processes that limit their practical applicability.

The advent of SSL has enabled the use of extensive unlabeled datasets to pre-train models that capture comprehensive utterance-level representations [12]. Yet, such representations may overlook critical phonetic or lexical subtleties that are pivotal for recognizing nuanced emotional expressions. Drawing inspiration from the Transformer TTS architecture [15], which efficiently leverages both phoneme and mel spectrogram inputs to generate high-fidelity audio, we propose to harness a similar mechanism for SER. In Transformer TTS, the integration of fine-grained phoneme sequences is crucial for synthesizing natural speech. Analogously, in SER, incorporating detailed phoneme-level information alongside word-level cues can enhance the recognition of stress and emphasis within spoken language.

To address these challenges, we propose the HarmoFusion framework—a novel multi-granularity strategy that fuses pre-trained global representations with fine-grained, localized features. HarmoFusion employs a dedicated transformer module to mediate cross-modal interactions among voice fragments, words, and phonemes. By exploring multiple fusion techniques (such as concatenation, element-wise multiplication, and attention-driven weighting), our approach is designed to capture both the overall sentiment and the subtle, context-dependent nuances of speech. A typical formulation within HarmoFusion for merging word-level (**W**) and phoneme-level (**P**) embeddings, which is further refined by a vanilla transformer layer to integrate sequential information robustly.

In summary, our contributions can be encapsulated as follows:

- We propose **HarmoFusion**, a unified framework that leverages multi-granularity fusion to integrate global and fine-grained features for multimodal emotion analysis.
- We introduce a hierarchical transformer-based module that effectively captures cross-modal interactions among voice fragments, words, and phonemes, thereby enhancing the overall feature representation.
- We provide extensive experimental validation on the IEMOCAP dataset [17], where HarmoFusion demonstrates significant improvements over state-of-the-art methods in terms of accuracy and robustness.
- We conduct comprehensive ablation studies and introduce additional loss formulations to quantify the contribution of fine-grained interactions in the overall performance.

This work not only advances the current state-of-the-art in SER but also sets a new direction for future research in the fusion of multimodal information using unified multi-granularity approaches.

2. Related Work

The evolution of Speech Emotion Recognition (SER) has traversed multiple paradigms. In the early days, classical machine learning techniques such as the Hidden Markov Model (HMM) [18] and the Gaussian Mixture Model (GMM) [19] were predominantly used. These models relied on handcrafted low-level acoustic features or high-level statistical descriptors to characterize speech signals. Although these approaches were pioneering, their dependence on manual feature engineering limited their ability to fully capture the intricate and dynamic nature of emotional expressions.

Subsequently, the advent of deep learning marked a significant turning point in SER research. Researchers began to harness the power of deep neural networks to learn hierarchical representations directly from raw input data. For instance, D. Bertero et al. [20] developed a convolutional neural network (CNN) architecture that extracted high-level features from raw spectrograms. The CNN's ability to learn local patterns in time-frequency representations allowed for a more robust encoding of emotional cues. In parallel, A. Satt et al. [21] proposed an end-to-end framework that combined CNNs with Long Short-Term Memory (LSTM) networks. This hybrid model was designed to capture both local feature patterns and long-term contextual dependencies, thus enhancing the system's capacity to model the temporal evolution of emotions.

With the increasing availability of multimodal data, the SER community shifted its focus towards integrating audio with textual information to exploit the complementary nature of these modalities. S. Yoon et al. [13] utilized Recurrent Neural Networks (RNNs) to encode both audio and text. Their approach involved using the final hidden state of one modality as a query in an attention mechanism, while the other modality provided the key-value pairs. Although this method successfully merged the two sources of information, the cross-modal interactions were only partially exploited, leaving room for improvement in capturing interdependencies.

In a similar vein, H. Xu et al. [14] advanced the field by designing a model that employed LSTM networks to learn the alignment between audio and text features through attention mechanisms. This strategy underscored the importance of temporal alignment in bridging the gap between different modalities. However, the focus was primarily on aligning cross-modal information, and less emphasis was placed on exploring the rich intra-modal dynamics within each individual stream.

Further progress was achieved by H. Li et al. [3], who introduced a fine-grained emotion recognition framework. Their method utilized a temporal mean-max alignment pooling strategy coupled with a cross-modality module to fuse aligned audio and text inputs. This approach enabled the model to capture subtle variations in emotional expression at a more granular level. Despite its effectiveness, the requirement for pre-aligned audio and text imposed additional preprocessing constraints, which could limit scalability and real-world applicability.

In recent years, the emergence of transformer architectures has provided a new perspective on SER. The self-attention mechanism has revolutionized many areas of machine learning by allowing models to capture global contextual relationships without the need for sequential processing. Although initially popularized in natural language processing, such attention mechanisms have been adapted to SER to capture both long-range dependencies and fine-grained interactions within speech data. This advancement has paved the way for hybrid models that combine the strengths of recurrent and attention-based approaches.

Moreover, recent efforts have concentrated on sophisticated multimodal fusion strategies. Beyond the conventional early and late fusion techniques, methods such as cross-modal attention and adaptive gating mechanisms have been proposed. One such formulation involves dynamically weighting different modalities based on their contribution to the emotion recognition task. Such dynamic fusion mechanisms allow the system to emphasize the most informative cues at each time step, thus enhancing overall performance.

Parallel to these developments, the integration of large-scale pre-trained models has shown promising results. Pre-trained models like BERT for textual data and Wav2Vec for speech have been successfully applied to SER, leveraging vast amounts of unlabeled data to generate robust utterance-level representations. However, while these global representations capture the overall sentiment, they often miss the localized, fine-grained emotional cues. Addressing this shortfall, our proposed HarmoFusion framework seeks to unify global pre-trained embeddings with localized features extracted at the phoneme and word levels. This dual-level approach not only harnesses the comprehensive contextual understanding of pre-trained models but also preserves the subtle details necessary for accurate emotion detection.

In addition, recent studies have explored strategies to mitigate the imbalance between modalities in SER. Techniques such as modality dropout, where one modality is intentionally omitted during training, force the model to develop more robust representations from the remaining data. Other approaches involve adversarial training methods designed to align the feature distributions across modalities, thus reducing modality-specific biases. These innovations highlight the ongoing challenge of effectively fusing heterogeneous data sources and underscore the need for more integrated approaches like HarmoFusion.

To summarize, the trajectory of SER research has evolved from relying on handcrafted features and classical statistical models to leveraging deep learning, attention mechanisms, and multimodal fusion techniques. Early methods provided valuable insights but were limited by their dependence on manual feature extraction. The transition to deep neural networks enabled more powerful, data-driven feature learning, while the subsequent adoption of attention and transformer-based models further refined the capture of temporal and contextual dependencies. Nonetheless, the challenge of integrating complementary information from disparate modalities persists. Our HarmoFusion framework addresses this challenge by harmoniously merging global and fine-grained representations, thereby offering a unified approach that improves both accuracy and robustness in multimodal emotion recognition.

Looking ahead, future research in this area may benefit from further exploration of cross-modal alignment techniques, the incorporation of additional modalities (such as visual cues), and the development of more adaptive fusion strategies. The continual evolution of SER methods promises to yield models that are not only more accurate but also more resilient in the face of real-world variability.

3. Proposed Methods and Implementation Details

In this section, we introduce our proposed methodology, which we term **HarmoFusion**. First, we detail the unified multi-granularity transformer framework that forms the core of HarmoFusion. Then, we describe the integration of pre-trained global representations with fine-grained features, culminating in a comprehensive multi-granularity model.

3.1. HarmoFusion: The Unified Multi-Grained Transformer Framework

This subsection elaborates on the design and components of HarmoFusion. Our framework is inspired by Transformer TTS [15], which builds upon the advances of Tacotron2 [22] and the transformer architecture [16]. However, in HarmoFusion we tailor the architecture specifically for the SER task by combining both phoneme-level and word-level information in a deep fusion strategy.

3.1.1. Framework Architecture

The HarmoFusion architecture begins by converting the input text into both phoneme and word sequences. Each token is then mapped into corresponding embedding spaces. The phoneme sequence is processed via a convolutional neural network (CNN) layer, where convolutional filters extract local phonetic features, and a subsequent max-pooling operation aggregates these features into fixed-length vectors. In parallel, the word sequence is embedded using pre-trained embeddings from Glove [26]. These embeddings serve as complementary sources of fine-grained and semantic information.

After obtaining the two sets of embeddings, the framework passes the concatenated representations through a highway network [23] to enable adaptive information flow. The transformation in the highway network is governed by

$$Z(u) = H(u) \cdot T(u) + u \cdot (1 - T(u)),$$

where u denotes the concatenated embedding vector, $H(u)$ is an affine transformation followed by a non-linear activation (e.g., ReLU [27]), and $T(u)$ is a gating function that learns to balance the transformed input with the original signal.

Subsequently, the output of the highway network is fed into an encoder pre-net, which consists of a three-layer CNN followed by a projection layer. This module serves to refine the sequential representation before entering the text encoder. In addition, the mel spectrogram is processed through a dedicated two-layer fully connected network to extract audio features.

A key innovation in HarmoFusion is the cross-modality interaction module. Here, the outputs from the text encoder and the processed mel spectrogram are combined using attention mechanisms.

In particular, we employ both self-attention and encoder-decoder attention layers to align and integrate features across modalities. A representative self-attention operation is defined as:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V,$$

where Q, K, and V are the query, key, and value matrices, respectively, and d_k is the dimensionality of the key vectors.

To further consolidate sequential dependencies, we incorporate additional vanilla transformer blocks. These blocks include position-wise feed-forward networks, whose operation can be expressed as:

$$\text{FFN}(x) = \max(0, xW_1 + b_1)W_2 + b_2,$$

where W_1, W_2, b_1, and b_2 are learnable parameters.

3.1.2. Overall Process and Inference Strategy

For the SER task, HarmoFusion leverages both the actual mel spectrogram and the textual input during both training and inference. Analogous to BERT's `[CLS]` token, we prepend a dummy mel vector to the mel spectrogram sequence:

$$m = (m_{\text{dummy}}, m_1, m_2, \ldots, m_{T'}),$$

where m_{dummy} is designed to capture a global summary of the spectrogram. During inference, rather than predicting the mel spectrogram in a sequential manner, the entire (golden) mel spectrogram is provided as input, ensuring that the emotion category probability is computed with full context. Formally, given a text sequence $x = (x_1, x_2, \ldots, x_T)$ and the mel spectrogram m, the probability p of an emotion category is calculated by the function:

$$p = g(x, m),$$

where $g(\cdot)$ encapsulates the HarmoFusion model's complex mapping from the joint text and audio features to the final emotion classification.

For comparison, in a typical TTS task the conditional probability of generating a mel spectrogram output o_t is given by:

$$f(o_t \mid x_1, \ldots, x_T) = f(o_t \mid o_{<t}, x),$$

where each output o_t is conditioned on all previous predictions. In contrast, our SER framework utilizes the entire mel spectrogram as input to yield a single, robust emotion prediction.

3.1.3. Text to Phoneme and Word Embedding Conversion

The conversion of text into multi-level representations is a cornerstone of HarmoFusion. Initially, the text is decomposed into its constituent phonemes using a phoneme extractor. This fine-grained representation is crucial because specific phonemes may carry significant emotional emphasis. Following the approach of [24,25], a CNN is applied to the phoneme sequence. The output of the CNN is then aggregated via max-pooling over the entire temporal dimension to generate a fixed-size vector for each word.

Simultaneously, the word-level semantics are captured using pre-trained Glove embeddings [26]. This dual representation ensures that both detailed acoustic features and high-level semantic cues are preserved.

3.1.4. Fusion of Phoneme and Word Embeddings

To effectively combine the fine-grained phoneme features with the robust semantic word features, we explore two primary fusion strategies. The first approach is simple concatenation, where the phoneme embedding vector and the word embedding vector are concatenated to form a composite representation:

$$u = [u_{\text{phoneme}}; u_{\text{word}}],$$

which is then passed through the encoder pre-net for further processing.

The second, more sophisticated approach employs a highway network to fuse the concatenated embeddings. This method allows the network to learn a dynamic balance between the transformed features and the original input. The fusion operation is defined as:

$$Z(u) = H(u) \cdot T(u) + u \cdot (1 - T(u)),$$

where $H(u) = \text{ReLU}(W_H u + b_H)$ is an affine transformation with ReLU activation, and $T(u) = \sigma(W_T u + b_T)$ is the gating function with a sigmoid activation $\sigma(\cdot)$. This mechanism not only facilitates information flow but also helps in mitigating vanishing gradient issues, ensuring effective multi-level fusion.

3.1.5. Transformer-Based Modules for Multi-Modal Integration

In HarmoFusion, the backbone of our framework is composed of transformer modules, which are utilized in several key stages:

- **Text Encoder:** Incorporates self-attention layers to merge information from both phoneme and word representations.
- **Cross-Modality Interaction Module:** Integrates features from the text encoder and the processed mel spectrogram using both self-attention and encoder-decoder attention layers. This module refines the joint representation by focusing on the interdependencies between modalities.
- **Deep Fusion Module:** Further aggregates the sequential multimodal representations using additional transformer blocks. The output corresponding to the dummy mel vector is extracted as the global feature, which is then projected linearly to generate the final logits.

An additional positional encoding is added to all transformer inputs to incorporate temporal information:

$$PE_{(pos,2i)} = \sin\left(\frac{pos}{10000^{2i/d_{\text{model}}}}\right), \quad PE_{(pos,2i+1)} = \cos\left(\frac{pos}{10000^{2i/d_{\text{model}}}}\right),$$

where pos represents the position and d_{model} is the dimensionality of the model.

3.1.6. Loss Function and Optimization

For the TTS auxiliary task, HarmoFusion generates both the mel spectrogram and a stop token, and these predictions are compared with the ground truth to compute the TTS loss. However, our primary objective is emotion classification. Although we experimented with multi-task learning approaches inspired by [28], we observed that joint optimization did not yield additional benefits in our scenario.

Therefore, we solely utilize the cross-entropy loss for the emotion classification head. The loss is defined as:

$$L = -\sum_{i=1}^{N}\sum_{k=1}^{K} y_{i,k} \log p_{i,k},$$

where $p_{i,k}$ is the predicted probability for class k for the i-th utterance, and $y_{i,k}$ is the ground truth indicator function. Additionally, to enforce consistency between intermediate representations, we introduce a regularization term:

$$\mathcal{L}_{\text{reg}} = \lambda \cdot \left\|\mathbf{F} - \mathbf{F}'\right\|^2,$$

where $\mathbf{F}$ and $\mathbf{F}'$ denote the feature representations from different layers of the deep fusion module, and λ is a weighting hyperparameter. The total loss is then given by:

$$\mathcal{L}_{\text{total}} = L + \mathcal{L}_{\text{reg}}.$$

3.2. Integration of Global and Fine-Grained Representations

3.2.1. Pre-trained and Multi-Level Components

Recent advances in natural language processing have been driven by large-scale pre-trained models. In our framework, we leverage the power of BERT, which comprises 12 layers with an embedding dimension of 768, to obtain robust, sentence-level representations. This pre-trained model has demonstrated state-of-the-art performance in numerous classification tasks, largely due to the efficacy of its `[CLS]` token, which encapsulates the overall semantic information of a sequence [12].

The fine-grained features extracted by the HarmoFusion framework (detailed in the previous subsection) complement these global representations. The deep fusion module in HarmoFusion produces a comprehensive representation that encapsulates the nuanced interactions between phoneme-level, word-level, and acoustic features.

3.2.2. Fusion Pipeline for Global and Fine-Grained Representations

Our multi-granularity model unifies the pre-trained BERT embeddings with the detailed representations generated by HarmoFusion. The overall pipeline is as follows:

1. The BERT model processes the input text and outputs a `[CLS]` embedding, denoted by $\mathbf{C} \in \mathbb{R}^{768}$, which represents the entire utterance.
2. Concurrently, the HarmoFusion module produces a fine-grained feature vector $\mathbf{F}_{\text{fine}}$ from the deep fusion module.
3. Both embeddings are projected into a common feature space via learnable projection matrices:

$$\tilde{\mathbf{C}} = W_C\mathbf{C} + b_C, \quad \tilde{\mathbf{F}}_{\text{fine}} = W_F\mathbf{F}_{\text{fine}} + b_F,$$

   where W_C, W_F and b_C, b_F are trainable parameters.
4. A late fusion strategy is employed by concatenating the projected vectors:

$$\mathbf{Z} = \left[\tilde{\mathbf{C}}; \tilde{\mathbf{F}}_{\text{fine}}\right].$$

5. The concatenated vector $\mathbf{Z}$ is then passed through a classification head consisting of a fully connected layer followed by a softmax activation to output the final emotion classification logits.

To further enhance the fusion process, we introduce an auxiliary attention mechanism that dynamically weighs the contributions of the global and fine-grained features:

$$\alpha = \sigma(W_\alpha\mathbf{Z} + b_\alpha),$$

where α is a scalar weight (or vector of weights) determining the relative importance of each modality, and $\sigma(\cdot)$ denotes the sigmoid function. The final representation used for classification is then computed as:

$$\mathbf{Z}_{\text{final}} = \alpha \cdot \tilde{\mathbf{C}} + (1 - \alpha) \cdot \tilde{\mathbf{F}}_{\text{fine}}.$$

This adaptive fusion strategy ensures that the model can balance global context and local details effectively, leading to improved performance in emotion recognition tasks.

In summary, the HarmoFusion framework integrates sophisticated transformer-based processing of multi-level textual and acoustic features with robust pre-trained representations, enabling a comprehensive analysis of emotional cues. The combination of advanced attention mechanisms, dynamic fusion strategies, and regularization techniques culminates in a system that is both accurate and resilient, paving the way for more effective multimodal emotion recognition.

4. Experiments

In this section, we detail the experimental settings, including the dataset description, implementation specifics, and comprehensive evaluation results of our proposed HarmoFusion framework. We also introduce additional technical formulas to quantify our evaluation metrics and provide an in-depth discussion of the results.

4.1. Dataset

We conduct our experiments on the IEMOCAP dataset [17], which is widely recognized in emotion recognition research. IEMOCAP comprises approximately 12 hours of audiovisual recordings, including video, speech, and text transcriptions. In our study, we solely utilize the audio signals and their corresponding transcriptions.

For consistency with previous studies [13], we focus on four emotion categories: angry (1103 utterances), sad (1084 utterances), neutral (1708 utterances), and happy (1636 utterances, with excited merged into happy), yielding a total of 5531 utterances. The dataset is partitioned into five folds with a split of 3:1:1 for training, development, and testing, respectively. To mitigate the impact of random initialization and sampling, each experiment is repeated three times and the average performance is reported.

For evaluating the multi-granularity model, we consider a more fine-grained categorization with six emotion classes: angry (1103), happy (595), excited (1041), sad (1084), frustrated (1849), and neutral (1708), resulting in 7380 samples. Following the protocol in [29], the dataset is divided into 70%, 10%, and 20% for training, development, and testing, respectively.

In order to rigorously assess the performance, we calculate the Weighted Accuracy (WA) and Unweighted Accuracy (UA) using the following formulas:

$$\mathrm{WA} = \frac{\sum_{i=1}^{K} n_i \cdot \mathrm{Accuracy}_i}{\sum_{i=1}^{K} n_i}, \quad \mathrm{UA} = \frac{1}{K}\sum_{i=1}^{K} \mathrm{Accuracy}_i,$$

where K is the total number of classes and n_i denotes the number of samples in class i.

4.2. Implementation Details

The proposed models are implemented using the PyTorch deep learning framework. For acoustic processing, we extract 128-dimensional filterbank features from speech signals with a window size of 25ms and a hop size of 12ms. To ensure robust feature representation, these filterbank features are normalized and fed into the network.

For textual input, we utilize 300-dimensional Glove embeddings [26]. The hidden size for all transformer layers is set to 128. All experiments are conducted on a Tesla V100 GPU. The Adam optimizer [30] is employed with a learning rate of 1×10^{-5} and a batch size of 4 for the single-modality HarmoFusion experiments. Early stopping is applied based on the WA on the validation set.

For the multi-granularity model, which integrates the HarmoFusion outputs with pre-trained BERT representations, the learning rate is increased to 5×10^{-5} and the batch size is set to 8, as guided by [29]. In this setting, the F1 score on the validation set is used for early stopping. For evaluation, both binary accuracy and F1 score are computed. The F1 score is calculated using:

$$F1 = 2 \cdot \frac{\mathrm{Precision} \times \mathrm{Recall}}{\mathrm{Precision} + \mathrm{Recall}},$$

and to ensure uniformity, the text length is capped at 50 tokens for the Albert-based experiments.

4.3. Multilevel Transformer Performance Evaluation

Table 1 summarizes the performance of various models on the IEMOCAP dataset under a 5-fold cross-validation setting. In addition to comparing against prior state-of-the-art approaches, we also conduct an extensive ablation study to assess the contributions of different components within HarmoFusion.

Table 1. Performance comparison between HarmoFusion and previous state-of-the-art approaches. The reported scores (WA and UA) are averaged over 5-fold cross validation with standard deviations.

Method	WA	UA
S. Yoon et al. [13]	0.685 ± 0.010	0.690 ± 0.011
H. Xu et al. [14]	0.688 ± 0.007	0.693 ± 0.006
H. Li et al. [3]	0.717 ± 0.003	0.725 ± 0.004
HarmoFusion (Proposed)	**0.740** ± **0.002**	**0.752** ± **0.002**
Ablation Study		
Phoneme only	0.686 ± 0.003	0.696 ± 0.005
Word only	0.715 ± 0.002	0.723 ± 0.003
Concatenation	0.737 ± 0.003	0.745 ± 0.003
Highway network	**0.740** ± **0.002**	**0.752** ± **0.002**
w/o Deep Fusion Module	0.732 ± 0.008	0.742 ± 0.007

As illustrated in Table 1, our HarmoFusion model achieves the highest WA and UA scores compared to previous methods. The ablation study further indicates that while both phoneme and word embeddings contribute positively to performance, their fusion via the highway network yields the best results. The removal of the deep fusion module leads to a noticeable performance drop, underscoring its critical role in capturing inter-modal interactions.

4.4. Transformer Module Configuration Analysis

To investigate the impact of transformer layer configurations on model performance, we conduct experiments by varying the number of layers in different modules of HarmoFusion. Table 2 presents the performance across several configurations for the text encoder, cross-modality interaction module, and deep fusion module.

Table 2. Effect of different transformer module layer configurations on performance. Each configuration is evaluated using WA and UA metrics. The best performance is obtained with 1-layer text encoder, 1-layer cross-modality interaction module, and 2-layer deep fusion module.

Text Encoder	Cross-Mod	Deep Fusion	WA	UA
3	3	1	0.724	0.735
2	2	1	0.728	0.737
1	1	1	0.729	0.740
1	1	2	**0.740**	**0.752**
1	1	3	0.728	0.738
2	2	2	0.735	0.744
2	2	3	0.725	0.732

The results in Table 2 demonstrate that a relatively shallow configuration, particularly with a 1-layer text encoder and cross-modality module combined with a 2-layer deep fusion module, yields optimal performance on the IEMOCAP dataset. This indicates that deeper models do not necessarily improve performance and that careful architectural tuning is essential for achieving superior results.

4.5. Multi-granularity Model Performance Evaluation

To further assess the benefits of integrating global pre-trained representations with the fine-grained features extracted by HarmoFusion, we compare the performance of the individual components with our combined multi-granularity approach. Table 3 summarizes these results.

Table 3. Performance comparison among different components. The multi-granularity model, which fuses pre-trained BERT embeddings with HarmoFusion outputs, outperforms both individual components in terms of WA and UA.

Component	WA	UA
BERT (Pre-trained)	0.700 ± 0.004	0.703 ± 0.002
HarmoFusion (Fine-grained)	0.740 ± 0.003	0.752 ± 0.002
Multi-granularity Model (Fusion)	**0.755 ± 0.003**	**0.760 ± 0.004**

As shown in Table 3, the multi-granularity model achieves a substantial improvement over the individual components. By fusing the global utterance-level information captured by BERT with the fine-grained details extracted by HarmoFusion, our approach leverages complementary strengths, leading to enhanced emotion recognition performance. The improvements in both WA and UA metrics highlight the effectiveness of our late fusion strategy in unifying multi-scale representations.

Overall, these experimental results validate the effectiveness of HarmoFusion and the multi-granularity fusion strategy. The rigorous evaluation on IEMOCAP demonstrates that our proposed models not only outperform previous state-of-the-art approaches but also benefit from a well-balanced integration of both global and local information.

5. Conclusions and Future Directions

In this paper, we introduced a novel framework named **HarmoFusion** that enables fine-grained interaction across different modalities, including voice fragments, words, and phonemes, for speech emotion recognition. Our approach represents a pioneering adaptation of the Transformer TTS structure to the SER task, demonstrating for the first time how such a model can be effectively repurposed to capture the nuanced emotional cues present in speech data. By seamlessly integrating detailed local representations with robust, pre-trained utterance-level features, HarmoFusion offers a comprehensive solution that bridges the gap between micro-level signal characteristics and macro-level semantic understanding.

Our extensive experimental evaluations, conducted on the IEMOCAP dataset, have shown that the HarmoFusion framework consistently outperforms existing state-of-the-art methods. The improvements in performance are attributed to the careful design of our multi-granularity fusion strategy, which effectively leverages both fine-grained details and high-level contextual information. This achievement is particularly noteworthy because it highlights the potential of combining complementary representations without the need for extensive alignment or manual intervention. The encouraging results suggest that our method can serve as a reliable reference for future research in emotion recognition, as well as in other domains that require the integration of heterogeneous data sources.

Beyond the technical contributions, our work also emphasizes the practical advantages of the proposed framework. HarmoFusion not only enhances the accuracy and robustness of emotion classification but also simplifies the model pipeline by reducing the reliance on complex pre-processing and manual alignment tasks. We are committed to open science and plan to make our code publicly available, which will facilitate replication of our results and promote further advancements in this area.

Looking forward, several exciting research directions emerge from our study. One promising avenue is the integration of additional acoustic pre-trained models, such as Wav2vec 2.0, to further enrich the representation of speech signals. Such integration could lead to even greater performance improvements by exploiting large-scale unlabeled data and the inherent structure of speech. Moreover, future work could explore the extension of HarmoFusion to incorporate visual and contextual cues, thereby creating a more holistic multimodal emotion recognition system. Another potential research direction involves the development of adaptive fusion mechanisms that dynamically adjust the contribution of each modality based on the specific characteristics of the input data. This adaptive approach may lead to models that are more resilient to variations in input quality and better suited to real-world applications.

In summary, our work on HarmoFusion represents a significant step forward in the field of speech emotion recognition. By unifying global and fine-grained representations within a single framework, we have demonstrated a new and effective strategy for capturing the complexities of human emotion. The success of HarmoFusion paves the way for future research that further explores and refines multi-granularity techniques, ultimately advancing the development of more intelligent and responsive human-computer interaction systems.

References

1. N. Sato and Y. Obuchi, "Emotion recognition using mel-frequency cepstral coefficients," *IMT*, vol. 2, no. 3, pp. 835–848, 2007.
2. Z. Zhao, Y. Wang, and Y. Wang, "Multi-level fusion of wav2vec 2.0 and BERT for multimodal emotion recognition," in *Interspeech*, 2022, pp. 4725–4729.
3. H. Li, W. Ding, Z. Wu, and Z. Liu, "Learning fine-grained cross modality excitement for speech emotion recognition," in *Interspeech*, 2021, pp. 3375–3379.
4. J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: pre-training of deep bidirectional transformers for language understanding," in *NAACL-HLT*, 2019, pp. 4171–4186.
5. Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut, "ALBERT: A lite BERT for self-supervised learning of language representations," in *ICLR*, 2020.
6. Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized BERT pretraining approach," *CoRR*, vol. abs/1907.11692, 2019.
7. S. Schneider, A. Baevski, R. Collobert, and M. Auli, "wav2vec: Unsupervised pre-training for speech recognition," in *Interspeech*, 2019, pp. 3465–3469.
8. A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," in *NeurIPS*, 2020.
9. A. Baevski, S. Schneider, and M. Auli, "vq-wav2vec: Self-supervised learning of discrete speech representations," in *ICLR*, 2020.
10. F. A. Acheampong, H. Nunoo-Mensah, and W. Chen, "Transformer models for text-based emotion detection: a review of bert-based approaches," *Artif. Intell. Rev.*, vol. 54, no. 8, pp. 5789–5829, 2021.
11. L. Pepino, P. Riera, and L. Ferrer, "Emotion recognition from speech using wav2vec 2.0 embeddings," in *Interspeech*, 2021, pp. 3400–3404.
12. S. Siriwardhana, A. Reis, R. Weerasekera, and S. Nanayakkara, "Jointly fine-tuning "bert-like" self supervised models to improve multimodal speech emotion recognition," in *Interspeech*, 2020, pp. 3755–3759.
13. S. Yoon, S. Byun, and K. Jung, "Multimodal speech emotion recognition using audio and text," in *SLT*, 2018, pp. 112–118.
14. H. Xu, H. Zhang, K. Han, Y. Wang, Y. Peng, and X. Li, "Learning alignment for multimodal emotion recognition from speech," in *Interspeech*, 2019, pp. 3569–3573.
15. N. Li, S. Liu, Y. Liu, S. Zhao, and M. Liu, "Neural speech synthesis with transformer network," in *IAAI*, 2019, pp. 6706–6713.
16. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *NIPS*, 2017, pp. 5998–6008.
17. C. Busso, M. Bulut, C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, "IEMOCAP: interactive emotional dyadic motion capture database," *Lang. Resour. Evaluation*, vol. 42, no. 4, pp. 335–359, 2008.
18. T. L. Nwe, S. W. Foo, and L. C. D. Silva, "Speech emotion recognition using hidden markov models," *Speech Commun.*, vol. 41, no. 4, pp. 603–623, 2003.
19. M. E. Ayadi, M. S. Kamel, and F. Karray, "Speech emotion recognition using gaussian mixture vector autoregressive models," in *ICASSP*, 2007, pp. 957–960.
20. D. Bertero and P. Fung, "A first look into a convolutional neural network for speech emotion detection," in *ICASSP*, 2017, pp. 5115–5119.
21. A. Satt, S. Rozenberg, and R. Hoory, "Efficient emotion recognition from speech using deep learning on spectrograms," in *Interspeech*, 2017, pp. 1089–1093.

doi:10.20944/preprints202502.1712.v1

22. J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Ryan, R. A. Saurous, Y. Agiomyrgiannakis, and Y. Wu, "Natural TTS synthesis by conditioning wavenet on MEL spectrogram predictions," in *ICASSP*, 2018, pp. 4779–4783.
23. R. K. Srivastava, K. Greff, and J. Schmidhuber, "Highway networks," *CoRR*, vol. abs/1505.00387, 2015.
24. M. J. Seo, A. Kembhavi, A. Farhadi, and H. Hajishirzi, "Bidirectional attention flow for machine comprehension," in *ICLR*, 2017.
25. Y. Kim, "Convolutional neural networks for sentence classification," in *EMNLP*, 2014, pp. 1746–1751.
26. J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *EMNLP*, 2014, pp. 1532–1543.
27. X. Glorot, A. Bordes, and Y. Bengio, "Deep sparse rectifier neural networks," in *AISTATS*, vol. 15, 2011, pp. 315–323.
28. X. Cai, J. Yuan, R. Zheng, L. Huang, and K. Church, "Speech emotion recognition with multi-task learning," in *Interspeech*, 2021, pp. 4508–4512.
29. W. Dai, S. Cahyawijaya, Z. Liu, and P. Fung, "Multimodal end-to-end sparse model for emotion recognition," in *NAACL-HLT*, 2021, pp. 5305–5316.
30. D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *ICLR*, 2015.
31. 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.
32. 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.
33. Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, may 2015. URL http://dx.doi.org/10.1038/nature14539.
34. 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/.
35. 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.
36. 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.
37. 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.
38. 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.
39. 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.
40. 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.
41. Ling Zhuang, Hao Fei, and Po Hu. Knowledge-enhanced event relation extraction via event ontology prompt. *Inf. Fusion*, 100:101919, 2023.
42. 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.
43. 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.
44. 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.

45. 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.
46. 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.
47. 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.
48. 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.
49. 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.
50. Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*, 2016.
51. 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.
52. 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.
53. 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.
54. 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. URL https://aclanthology.org/N19-1423.
55. Shengqiong Wu, Hao Fei, Leigang Qu, Wei Ji, and Tat-Seng Chua. Next-gpt: Any-to-any multimodal llm. *CoRR*, abs/2309.05519, 2023.
56. 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.
57. 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.
58. Naman Jain, Pranjali Jain, Pratik Kayal, Jayakrishna Sahit, Soham Pachpande, Jayesh Choudhari, et al. Agribot: agriculture-specific question answer system. *IndiaRxiv*, 2019.
59. 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.
60. 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.
61. 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.
62. 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.
63. 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.
64. 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.

doi:10.20944/preprints202502.1712.v1

65. 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.
66. Sanjeev Arora, Yingyu Liang, and Tengyu Ma. A simple but tough-to-beat baseline for sentence embeddings. In *ICLR*, 2017.
67. 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.
68. 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.
69. 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.
70. 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.
71. 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.
72. 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.
73. 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.
74. 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.
75. 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.
76. 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.
77. 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.
78. 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.
79. 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.
80. 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.
81. Yu Zhao, Hao Fei, Wei Ji, Jianguo Wei, Meishan Zhang, Min Zhang, and Tat-Seng Chua. Generating visual spatial description via holistic 3D scene understanding. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7960–7977, 2023.

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.