Article

Not peer-reviewed version

Enhanced Modal Fusion Learning for Multimodal Sentiment Interpretation

Kayla Robinson , Ava Martinez , Ethan Turner *

Posted Date: 24 September 2024

doi: 10.20944/preprints202409.1887.v1

Keywords: sentiment analysis; robust learning; cross-modal integration

 doi:10.20944/preprints202409.1887.v

Article

Enhanced Modal Fusion Learning for Multimodal Sentiment Interpretation

Kayla Robinson, Ava Martinez and Ethan Turner *

University of Central Oklahoma
* Correspondence: ethan.turner@uco.edu

Abstract: Multimodal sentiment analysis is rapidly gaining traction due to its ability to comprehensively interpret opinions expressed in video content, which is ubiquitous across various digital platforms. Despite its promising potential, the field is hindered by the limited availability of high-quality, annotated datasets, which poses substantial challenges to the generalizability of predictive models. Models trained on such scarce data often inadvertently assign excessive importance to irrelevant features, such as personal attributes (e.g., eyewear), thereby diminishing their accuracy and robustness. To address this issue, we propose an Enhanced Modal Fusion Learning (EMFL) methodology aimed at significantly improving the generalization capabilities of neural networks. EMFL achieves this by optimizing the integration and interpretation processes of multimodal data, ensuring that sentiment-relevant features are prioritized over confounding attributes. Through extensive experiments conducted on multiple benchmark datasets, we demonstrate that EMFL consistently elevates the accuracy of sentiment predictions across verbal, acoustic, and visual modalities. These findings underscore EMFL's efficacy in mitigating the impact of non-relevant features and enhancing the overall performance of multimodal sentiment analysis models.

Keywords: sentiment analysis; robust learning; cross-modal integration

1. Introduction

Sentiment analysis traditionally focuses on text to determine the underlying emotional tone as positive, negative, or neutral [1,2,42]. With the advent of multimedia sharing platforms, there is an upsurge in video content featuring expressive modalities beyond text—spoken words [36,43], vocal dynamics, and visual cues. Multimodal sentiment analysis seeks to combine these elements to provide a more comprehensive understanding of sentiments expressed in videos [3,4].

Multimodal sentiment analysis is an increasingly crucial area of study within the domain of artificial intelligence, focusing on the interpretation and analysis of emotions from multiple types of data, such as text, audio, and video. This multidisciplinary field leverages advancements from several areas including natural language processing, computer vision, and audio analysis, to understand the nuances of human sentiment more comprehensively. The traditional approach to sentiment analysis primarily dealt with text. However, with the growth of video content on platforms like YouTube, Instagram, and TikTok, the need to analyze sentiment from not only text but also audio and visual cues has become essential. Each mode—text, audio, visual—can provide complementary information which, when integrated, offers a fuller understanding of the communicator's emotional state [1,5,6].

While the multimodal approach offers depth, the annotated datasets available for training are modest in size and variability [1,5,6]. This limitation often causes models to associate sentiment with incidental attributes of speakers, such as wearing glasses, rather than the intended sentiment indicators [49,50]. For instance, a scenario where models trained on sparse data associate negative sentiment with visual indicators such as eyewear illustrates the challenge of dataset bias. These biases act as confounding factors, statistically skewing results and reducing the efficacy of sentiment analysis models across different datasets [7,8]. An examination of the MOSI dataset revealed significant dependencies between sentiments expressed and individual identities, indicating potential biases in model training that could lead to erroneous sentiment predictions based on personal characteristics rather than actual sentiment [6,58].

 doi:10.20944/preprints202409.1887.v

One of the major challenges in multimodal sentiment analysis is the fusion of different modalities. How best to integrate these sources of data remains a significant research question. Early fusion and late fusion are common strategies [25,27]; the former integrates features at the beginning of the process [59,60], while the latter combines the outputs near the end. More sophisticated approaches, like hybrid fusion, attempt to capture the strengths of both methods [11]. Another challenge is the scarcity of labeled multimodal datasets, which are crucial for training machine learning models. The cost of obtaining high-quality annotations across multiple modalities can be prohibitive, limiting the availability of large-scale datasets [15].

To address these challenges, this paper proposes the Enhanced Modal Fusion Learning (EMFL) approach, a robust framework designed for neural networks, particularly convolutional neural networks. EMFL involves a dual-phase process where the first phase (*Selection*) identifies and isolates confounding factors, and the second phase (*Addition*) minimizes their impact by integrating stochastic disturbances into the data representation. This process ensures that the learning mechanism prioritizes relevant sentiment indicators over misleading biases.

We extensively evaluate the effectiveness of EMFL through a series of experiments conducted in a person-independent manner, ensuring that no individual appears in both training and testing datasets. Our findings confirm that EMFL not only improves within-dataset accuracy but also ensures robust generalization across multiple external datasets.

2. Related Work

Research in multimodal data utilization spans a broad spectrum of applications aimed at enhancing our understanding and interaction with human behavioral patterns. These applications include, but are not limited to, person detection and identification [9,10,62], human action recognition [11,12,64], and face recognition [13,14]. Each of these fields contributes to the foundational technologies essential for advanced sentiment analysis systems.

Sentiment analysis initially focused on textual data, exploring various levels of granularity from words [15,73] to phrases [16], sentences [17], and entire documents [2,70]. The evolution of this field saw the integration of deep learning models, notably recursive neural networks, which significantly enhanced the ability to understand complex emotional expressions in text.

Recent advancements in multimodal sentiment analysis emphasize the importance of effective modality integration techniques to improve the accuracy and robustness of sentiment detection. Researchers have explored various approaches to modality fusion, which include feature-level fusion, decision-level fusion, and hybrid approaches that aim to leverage the unique advantages of each method [12,51]. For instance, feature-level fusion involves combining features from different modalities before inputting them into the classifier, enhancing the model's ability to capture interdependencies between modalities [14]. On the other hand, decision-level fusion integrates the outputs of separate classifiers for each modality at a later stage, which has been shown to increase the system's resilience to errors in individual modal predictors [17].

Moreover, the development of hybrid fusion techniques has seen significant interest, combining both feature and decision-level methods to optimize performance. These approaches typically employ machine learning techniques such as Support Vector Machines (SVM) and neural networks to dynamically weigh the contribution of each modality based on its reliability and relevance to the sentiment analysis task [23]. This dynamic weighting system can adapt to the context of the data, potentially offering superior performance over static fusion methods.

In the realm of audio data, traditional methods often involve transcribing spoken content to text, followed by sentiment analysis [18]. Additionally, there is a growing interest in directly assessing the emotional state from vocal characteristics without the need for transcription [19]. This direct approach aligns with developments in recognizing emotional states through the Facial Action Coding System, which has been foundational for analyzing facial expressions [20,81]. The advent of convolutional

 doi:10.20944/preprints202409.1887.v

neural networks has further revolutionized this area by pinpointing affective regions in images to assess sentiments [21].

The integration of textual, acoustic, and visual data—often referred to as multimodal sentiment analysis—has recently seen significant interest and development [1]. Various innovative methods have been explored, ranging from simple data concatenation to sophisticated hybrid models that intelligently fuse modalities to predict sentiments more accurately [22–24]. Among these, convolutional neural networks stand out for their effectiveness in handling multimodal data fusion, achieving state-of-the-art results [25].

Building on this body of work, we propose the Enhanced Modal Fusion Learning (EMFL) methodology, which significantly improves the generalizability of neural networks across different data modalities. Our approach is designed to address the inherent challenges of multimodal sentiment analysis, such as modal disparity and data sparsity. We present extensive experimental results that demonstrate the superiority of EMFL in enhancing prediction accuracy not only within individual modalities—verbal, acoustic, and visual—but also in their integrated form. The next section details the EMFL methodology, focusing on its novel selection and addition phases designed to mitigate bias and enhance data interoperability.

3. Enhanced Multimodal Feature Learning (EMFL)

The primary objective of our research is to enhance the generalizability of multimodal sentiment analysis models by encouraging the model to prioritize sentiment-related features (e.g., individuals smiling when conveying positive emotions) over identity-specific attributes (e.g., wearing glasses). By doing so, we aim to mitigate the influence of confounding factors that may otherwise bias the sentiment prediction.

We formalize this challenge by introducing an input feature matrix X of dimensions $n \times p$, where n represents the number of utterances and p denotes the total number of features extracted from verbal, acoustic, and visual modalities. Additionally, we define a sentiment vector y of size $n \times 1$ corresponding to the sentiment label of each utterance. To account for speaker identities, we introduce a one-hot encoded matrix Z of size $n \times m$, where m is the number of unique individuals in the dataset.

Our proposed EMFL framework is designed to augment an existing (pre-trained) discriminative neural network, enhancing its robustness against confounding variables. To formally describe EMFL, we identify two fundamental components commonly found in discriminative neural network classifiers (such as Convolutional Neural Networks, CNNs): a representation learning component and a classification component.

For simplicity, let us denote the representation learning component by $g(\cdot;\theta)$, where θ signifies its parameters. We hypothesize that confounding factors are confined to a subset of the dimensions within $g(\cdot;\theta)$. Similarly, the classification component is denoted by $f(\cdot;\phi)$, with ϕ representing its parameters. Consequently, the complete neural network classifier can be expressed as $f(g(\cdot;\theta);\phi)$.

Within the EMFL framework, the representation learner $g(\cdot;\theta)$ captures identity-related features, which we term as *identity-related confounding dimensions*. The EMFL approach involves two main steps: firstly, identifying these confounding dimensions, and secondly, diminishing their impact by introducing noise.

To accurately select the *identity-related confounding dimensions*, EMFL incorporates a straightforward neural network component denoted by $h(\cdot;\delta)$, where δ represents its parameters. This component is tasked with predicting the *identity-related confounding dimensions* based on individual identities Z, by minimizing the discrepancy between $h(Z;\delta)$ and $g(X;\theta)$. As a result, $h(Z;\delta)$ effectively isolates the *identity-related confounding dimensions* within $g(X;\theta)$.

To compel the model to disregard the *identity-related confounding dimensions*, EMFL introduces Gaussian noise to these specific dimensions while simultaneously minimizing the prediction error. This ensures that the classification component $f(\cdot;\phi)$ learns to focus on the relevant representations,

 doi:10.20944/preprints202409.1887.v

effectively ignoring the noised confounding dimensions. The noise addition is facilitated through a Gaussian Sampling Layer as described in [26].

The uppermost section represents the Decision Making Layers (DML, denoted as $f(\cdot;\phi)$ in equations), responsible for making the final prediction. In our experiments, the DML is structured as a single-layer traditional neural network followed by a Logistic Regression Layer. The Gaussian Sampling Layer (GSL) assists the DML in fine-tuning parameters to differentiate between relevant and confounding representations. The lower left section, highlighted in purple, performs Mixed Representation Learning (MRL, denoted as $g(\cdot;\theta)$ in equations) and can utilize any representation learner such as CNNs, autoencoders, LSTMs, or even pre-trained models. In our experiments, we employ a state-of-the-art CNN as outlined in [25]. The lower right section, marked in red, undertakes Confounding Representation Learning (CRL, denoted as $h(\cdot;\delta)$ in equations). For our experiments, the CRL is implemented using a single-layer traditional neural network.

The Gaussian Sampling Layer integrates EMFL into the original neural network architecture through a Gaussian sampling process, where the mean is determined by the representations learned from the original bottom layers, and the variance is dictated by the representations learned from EMFL.

3.1. Gaussian Sampling Layer

The Gaussian Sampling Layer enhances a conventional neural network layer by incorporating a variance term, as discussed in [26]. Unlike a traditional neural network layer, whose output T is a deterministic function of the input M:

$$T = \psi(MW + B)$$

where W, B, and $\psi(\cdot)$ denote the weights, bias, and activation function respectively, the Gaussian Sampling Layer produces its output by sampling from a Gaussian distribution defined by the deterministic features input M and the variance features input Σ:

$$G \sim \mathcal{N}(M, \mathrm{diag}(\Sigma\Sigma^\top))$$

For brevity, we denote $\mathrm{diag}(\Sigma\Sigma^\top)$ as $d(\Sigma)$ in the subsequent discussions.

With EMFL integrated into the original model architecture, the ultimate objective of parameter learning is to optimize the following function:

$$\underset{\phi,\theta,\delta}{\operatorname{argmin}} \frac{1}{2}(y - f(g(X,\theta) + h(Z,\delta)\epsilon, \phi))^2 \tag{1}$$

where $\epsilon \sim \mathcal{N}(0,\sigma)$. This optimization problem is inherently non-convex and poses significant challenges even when $f(\cdot,\phi)$, $g(\cdot,\theta)$, and $h(\cdot,\delta)$ are linear functions. To address this complexity, we propose a heuristic algorithm named Enhanced Multimodal Feature Learning (EMFL), which efficiently navigates the search space for θ, δ, and ϕ in an iterative manner to achieve the optimization goal.

3.2. Enhanced Multimodal Feature Learning Algorithm

A fundamental prerequisite for our Enhanced Multimodal Feature Learning (EMFL) approach is the initial training of a discriminative neural classifier. In our experimental setup, this is accomplished by minimizing the following loss function:

$$\underset{\phi,\theta}{\operatorname{argmin}} \frac{1}{2}(y - f(g(X;\theta);\phi))^2$$

This loss function is a standard choice in discriminative neural networks, as evidenced in [27].

 doi:10.20944/preprints202409.1887.v

Our EMFL methodology is versatile and can be seamlessly integrated into various deep learning models, provided that the model can be decomposed into the classification component $f(\cdot, \phi)$ and the representation learning component $g(\cdot, \theta)$. Moreover, EMFL can be applied to pre-existing published models to enhance their performance without necessitating significant architectural changes.

The EMFL algorithm is structured into three distinct phases: 1) Representation Learning Phase, 2) Confounding Representation Matching Phase, and 3) Confounding Representation Elimination Phase.

These phases are delineated based on the specific representations they target. The overarching aim is to optimize the following objective:

$$\underset{\phi,\theta}{\operatorname{argmin}} \frac{1}{2}(y - f(g(X, \theta), \phi))^2 \tag{2}$$

This objective mirrors that of a traditional neural network. However, when combined with the optimization of additional functions, our EMFL algorithm significantly enhances the generalization capabilities of both $f(\cdot, \phi)$ and $g(\cdot, \theta)$, thereby enabling superior performance on unseen test data.

The Gaussian Sampling Layer (GSL) is utilized exclusively during the training phase, as its variance term is derived from the confounding representation. During the validation and testing phases, since identity information Z is unavailable, Z is represented as a vector of zeros, resulting in the output $t = \mu$. The subsequent layers operate as detailed below.

The lower left section comprises Mixed Representation Learning (MRL) models, which process the input X to generate a comprehensive mixed representation. The representation learners can be CNNs, autoencoders, or other suitable architectures. In our experiments, we adopt a state-of-the-art CNN configuration as described in [25]. The feature set X is consistently available across the training, validation, and testing phases.

Conversely, the lower right section consists of Confounding Representation Learning (CRL) models, which utilize the input Z to extract confounding representations. In our experiments, the CRL is implemented using a single-layer traditional neural network. It is important to note that Z is only accessible during the training phase as a set of one-hot vectors encoding individual identities. During validation and testing phases, Z is a vector of zeros.

3.2.1. Representation Learning Phase

The Representation Learning Phase focuses on fine-tuning the parameters of the Mixed Representation Learning (MRL) component to solve Equation 2. In this phase, only the input feature matrix X is utilized, and the representation is computed as $G = g(X; \theta)$. This phase is analogous to training a standard sentiment analysis deep learning model without considering confounding factors.

The figure illustrates the Decision Making Layers (DML) alongside their associated weights (represented as lines) and the intermediate representations (depicted as squares) that serve as inputs. The representation fed into the DML is a composite of both relevant (blue squares) and confounding (red squares) representations. At this stage, the DML lacks the capability to differentiate between these two types of representations, resulting in all its weights remaining active and contributing equally to the final prediction.

This phase is primarily intended to adjust the parameters of the MRL and DML components. Consequently, if pre-trained models are available, this phase can be omitted to expedite the training process.

3.2.2. Confounding Representation Matching Phase

Upon establishing the initial representation $g(X; \theta)$, the Confounding Representation Matching Phase seeks to identify and isolate the *identity-related confounding dimensions*. This is achieved by optimizing a new loss function specifically designed to uncover these dimensions. The optimization is carried out by tuning the parameters δ using the following loss function:

 doi:10.20944/preprints202409.1887.v

$$\arg\min_{\delta} \frac{1}{2}(g(X;\theta) - h(Z;\delta))^2 + \lambda\|\delta\|_1 \tag{3}$$

Here, λ is a hyperparameter that controls the strength of the sparsity regularization imposed on δ. This regularization is crucial to prevent overfitting, especially given that the output dimension of $h(\cdot;\delta)$ typically exceeds that of its input dimension.

During this phase, both X and Z are available as inputs. However, only the parameters δ of the confounding representation learner $h(\cdot;\delta)$ are updated.

The objective of this phase is to selectively identify the *identity-related confounding dimensions* within the original representation $g(X;\theta)$. By minimizing the difference between $g(X;\theta)$ and $h(Z;\delta)$, and given that Z encapsulates solely identity information, the optimization ensures that $h(Z;\delta)$ aligns with the identity-related components of $g(X;\theta)$. The L1 regularization term $\lambda\|\delta\|_1$ promotes sparsity in δ, ensuring that only the most pertinent dimensions are selected as confounders. The original model's weights (represented by the purple circle) remain fully active and interconnected with every dimension. In contrast, only a subset of weights within $h(\cdot;\delta)$ (depicted by the red circle) are active, corresponding to the identified *identity-related confounding dimensions*.

3.2.3. Confounding Representation Elimination Phase

Following the successful identification of the *identity-related confounding dimensions* in the Confounding Representation Matching Phase, the next step is to train a new neural network classifier that effectively "masks" these confounding dimensions. This is accomplished by introducing Gaussian noise to the identified dimensions, thereby rendering them non-informative. The classifier is then trained to focus on the remaining, relevant dimensions.

The Confounding Representation Elimination Phase is governed by the following loss function:

$$\operatorname{argmin}_{\phi} \frac{1}{2}(y - f(g(X;\theta) + h(Z;\delta) \circ \epsilon;\phi))^2 \tag{4}$$

where $\epsilon \sim \mathcal{N}(0, \sigma I)$ and $\circ$ denotes the element-wise product.

In this phase, only the parameters ϕ of the classification component $f(\cdot;\phi)$ are updated. The input to $f(\cdot;\phi)$ now comprises the original representation $g(X;\theta)$ augmented with the noised confounding representation $h(Z;\delta) \circ \epsilon$. The introduction of noise ensures that the *identity-related confounding dimensions* do not carry meaningful information, compelling the classifier to learn to ignore these dimensions and rely solely on the relevant features for accurate sentiment prediction. The *identity-related confounding dimensions* are contaminated with Gaussian noise. This contamination forces the model to disregard these non-informative dimensions, thereby allowing the weights to be optimized to concentrate on the remaining, informative dimensions.

It is important to note that EMFL introduces an additional set of parameters (δ) to be trained and requires the selection of two hyperparameters: λ in Equation 3 and σ in Equation 4. Strategies for selecting optimal values for λ and σ are elaborated upon in the supplementary material.

3.3. Algorithmic Workflow of EMFL

The EMFL algorithm operates through an iterative process encompassing the three aforementioned phases. The workflow is summarized as follows:

1. **Initialization**: Begin with a pre-trained discriminative neural network, comprising the representation learner $g(\cdot;\theta)$ and the classifier $f(\cdot;\phi)$.
2. **Representation Learning Phase**: Fine-tune the parameters θ and ϕ by minimizing the primary loss function (Equation 2), thereby optimizing the mixed representation learning without considering confounders.

 doi:10.20944/preprints202409.1887.v

3. **Confounding Representation Matching Phase**: Introduce the confounding representation learner $h(\cdot;\delta)$ and optimize δ by minimizing the loss function in Equation 3. This phase isolates the identity-related confounding dimensions within the representation.
4. **Confounding Representation Elimination Phase**: Incorporate Gaussian noise into the identified confounding dimensions and retrain the classifier $f(\cdot;\phi)$ by minimizing the loss function in Equation 4. This step ensures that the classifier learns to ignore the noised confounders.
5. **Iteration**: Repeat steps 2 through 4 until convergence is achieved, i.e., when further iterations do not yield significant improvements in performance.
6. **Final Model**: The resultant model, now robust against confounding factors, is evaluated on unseen test data to assess its generalization capabilities.

Through this iterative process, EMFL effectively disentangles relevant sentiment-associated features from identity-related confounders, thereby enhancing the model's ability to generalize across diverse datasets and scenarios.

3.4. Advantages of EMFL

The Enhanced Multimodal Feature Learning (EMFL) framework offers several key advantages:

- **Robustness to Confounders**: By explicitly identifying and mitigating the influence of identity-related confounding dimensions, EMFL ensures that the sentiment predictions are not biased by irrelevant identity features.
- **Versatility**: EMFL is model-agnostic and can be integrated into a wide range of deep learning architectures, including CNNs, LSTMs, and autoencoders, as well as pre-trained models, thereby broadening its applicability.
- **Improved Generalization**: By focusing on sentiment-relevant features, EMFL enhances the model's ability to generalize to new, unseen data, thereby improving overall prediction performance.
- **Scalability**: The framework is scalable to large datasets with numerous features and identities, making it suitable for real-world applications where data complexity is high.
- **Ease of Integration**: The three-phase approach of EMFL allows for straightforward integration into existing training pipelines without necessitating significant architectural modifications.

These advantages make EMFL a compelling approach for advancing the state-of-the-art in multimodal sentiment analysis and potentially other domains where confounding factors may impede model performance.

3.5. Implementation Considerations

When implementing the EMFL framework, several considerations must be addressed to ensure optimal performance:

- **Hyperparameter Tuning**: The selection of appropriate values for the hyperparameters λ and σ is crucial. These parameters control the sparsity of the confounding dimension selection and the magnitude of the introduced noise, respectively. Techniques such as cross-validation and grid search can be employed to identify optimal values.
- **Computational Overhead**: Introducing additional components such as the confounding representation learner $h(\cdot;\delta)$ and the Gaussian Sampling Layer (GSL) may increase computational requirements. Efficient implementation strategies and hardware acceleration can mitigate potential performance bottlenecks.
- **Data Quality and Representation**: The effectiveness of EMFL is contingent upon the quality and representativeness of the input features. Ensuring comprehensive and relevant feature extraction across all modalities is essential for the successful identification of confounding dimensions.

 doi:10.20944/preprints202409.1887.v

- **Scalability to Multiple Confounders**: While EMFL is designed to handle identity-related confounders, extending the framework to account for multiple types of confounders may require additional modifications and complexity.
- **Evaluation Metrics**: Employing appropriate evaluation metrics that accurately reflect the model's ability to generalize and its robustness to confounders is vital. Metrics such as accuracy, F1-score, and area under the ROC curve (AUC) can provide comprehensive insights into model performance.

Addressing these considerations during implementation will enhance the efficacy and reliability of the EMFL framework in practical applications.

4. Experiments

In this section, we conduct a comprehensive series of experiments across three distinct datasets to evaluate the efficacy of our proposed Enhanced Multimodal Feature Learning (EMFL) framework in enhancing the generalizability of discriminative neural classifiers. The primary metric for assessing generalizability involves cross-dataset validation, wherein models trained on one dataset are exclusively tested on the remaining two datasets. This approach ensures that the models are evaluated on unseen data distributions, thereby providing a robust measure of their generalization capabilities. All experiments adhere to a person-independent methodology, ensuring that no individual present in the training data appears in the test datasets, thereby eliminating potential data leakage and bias.

Table 1. Classification accuracy across three datasets for CNN and EMFL-CNN models across various modalities and their multimodal integrations. Models are initially trained and validated on a subset of 62 individuals from the MOSI dataset, and subsequently evaluated on two distinct test datasets comprising 31 individuals from MOSI, 47 individuals from YouTube, and 55 individuals from MOUD respectively.

		Within Dataset		Across Datasets			
		MOSI		YouTube		MOUD	
		CNN	EMFL-CNN	CNN	EMFL-CNN	CNN	EMFL-CNN
Single Modality	Text	0.678	**0.732**	0.605	**0.657**	0.522	**0.569**
	Audio	0.588	**0.618**	0.441	**0.564**	0.455	**0.549**
	Video	0.572	**0.636**	0.492	**0.549**	**0.555**	0.548
Double Modalities	Text+Audio	0.687	**0.725**	0.642	**0.652**	0.515	**0.574**
	Text+Video	0.706	**0.73**	0.642	**0.667**	0.542	**0.574**
	Audio+Video	**0.661**	0.621	0.452	**0.559**	0.533	**0.554**
All Modalities		0.715	**0.73**	0.611	**0.667**	0.531	**0.574**

4.1. Model Architectures

To thoroughly assess the performance improvements introduced by the EMFL framework, we compare the following models:

CNN: We utilize a state-of-the-art seven-layer Convolutional Neural Network (CNN) architecture, previously established for multimodal sentiment analysis [25]. This CNN serves as our baseline model, leveraging its proven efficacy in capturing intricate patterns across multiple data modalities.

EMFL-CNN: Building upon the baseline CNN, we integrate the EMFL framework to enhance its generalizability. After the CNN is fully trained, EMFL is applied to mitigate the influence of confounding factors, thereby refining the model's ability to focus on sentiment-relevant features. The confounding representation learner $h(\cdot;\delta)$ within EMFL is implemented as a neural perceptron [27], ensuring effective identification and suppression of identity-related confounders.

4.2. Datasets Utilized

Our experimental evaluation is conducted on three prominent multimodal sentiment analysis datasets:

 doi:10.20944/preprints202409.1887.v

MOSI (Multimodal Opinion-level Sentiment Intensity): Comprising 93 YouTube videos, the MOSI dataset encapsulates opinions from 93 unique individuals. It contains 2,199 utterances, each manually segmented and annotated for sentiment intensity [6]. The dataset is well-suited for sentiment analysis tasks due to its diverse expressions and comprehensive annotations.

YouTube: This dataset consists of 47 opinionated YouTube videos, encompassing a total of 280 utterances. Each video reflects the sentiments of one distinct individual, with sentiments manually annotated [1]. The YouTube dataset is characterized by its varying recording qualities and processing methodologies, providing a challenging testbed for cross-dataset generalization.

MOUD (Multimodal Opinion Utterance Dataset): The MOUD dataset includes 498 Spanish-language opinion utterances sourced from 55 unique individuals [5]. This dataset introduces an additional layer of complexity due to the necessity of translating Spanish transcripts into English, thereby testing the robustness of models across linguistic variations.

Despite originating primarily from the YouTube platform, these datasets exhibit significant differences in recording quality and post-curation processing steps. Furthermore, verbal features are extracted using disparate Automatic Speech Recognition (ASR) tools across the datasets. The MOUD dataset's verbal features require an extra translation step from Spanish to English, introducing potential variability in linguistic representation. These distinctions render the three datasets ideal candidates for evaluating the cross-dataset generalization capabilities of our models.

4.3. Feature Extraction Techniques

For each dataset, we employed a meticulous feature extraction process tailored to capture the nuances across textual, acoustic, and visual modalities:

Textual Features: We extracted word embeddings using a pre-trained Word2Vec model on the Google News corpus [28]. To enrich the semantic representation, a 6-dimensional binary vector indicating the Part-of-Speech (POS) tags—noun, verb, adjective, adverb, preposition, conjunction—was appended to each word embedding. Each utterance's text feature was constructed by concatenating the embeddings of all constituent words, padding with zeros to ensure uniform dimensionality. We set the maximum utterance length to 60 words, discarding any additional words beyond this limit, which accounted for only approximately 0.5% of utterances across datasets. The penultimate fully connected layer of the CNN was extracted to serve as the final textual input for training.

For the YouTube dataset, transcripts were generated using IBM Bluemix's Speech-to-Text API[1]. In contrast, the MOUD dataset required translation of Spanish transcripts into English to maintain consistency across all datasets.

Acoustic Features: Utilizing the openSMILE toolkit [29], we extracted low-level audio descriptors for each utterance. These descriptors included Mel-Frequency Cepstral Coefficients (MFCCs), pitch, and various voice quality metrics. Each utterance was segmented into 50 trunks, and the features within each trunk were averaged, resulting in a comprehensive 1,950-dimensional feature vector per utterance.

Visual Features: Each video frame was processed to extract facial characteristic points using the CLM-Z library [30]. Frames corresponding to specific utterances were identified through audio-visual synchrony. Each utterance was divided into 5 trunks, with features within each trunk averaged to produce a 2,075-dimensional visual feature vector per utterance.

This multi-faceted feature extraction approach ensures that the models are equipped with rich and diverse representations, capturing the essential aspects of sentiment expression across modalities.

---

[1] https://www.ibm.com/watson/developercloud/speech-to-text.html

 doi:10.20944/preprints202409.1887.v

4.4. Experimental Setup

To emulate real-world scenarios and rigorously evaluate the generalization performance of our models, we employed a stringent experimental setup:

Dataset Partitioning: We selected the first 62 individuals from the MOSI dataset to constitute the training and validation sets, encompassing a total of approximately 1,250 utterances. These utterances were randomly shuffled and partitioned into 1,000 training cases and 250 validation cases, maintaining an 80-20 split. This partitioning ensures that the model is trained and validated on a substantial and diverse subset of the data.

Test Sets: Three distinct test sets were constructed to evaluate cross-dataset performance:

1. **MOSI Test Set**: Comprising 546 utterances from the remaining 31 individuals in the MOSI dataset.
2. **YouTube Test Set**: Consisting of 195 utterances from 47 unique individuals in the YouTube dataset.
3. **MOUD Test Set**: Encompassing 450 utterances from 55 individuals in the MOUD dataset.

This partitioning ensures that the test sets are entirely disjoint from the training and validation sets, thereby providing an unbiased assessment of the models' generalization capabilities.

Training Protocol: The CNN model was initially trained on the training set and validated on the validation set. The model achieving the highest validation accuracy was selected for subsequent evaluations. Following this, the EMFL framework was applied to the pre-trained CNN to produce the EMFL-CNN model. This model underwent further training to minimize the validation loss, ensuring optimal performance.

Evaluation Metrics: Classification accuracy was employed as the primary metric for evaluating model performance across different modalities and datasets. Additionally, statistical significance of performance improvements was assessed using permutation tests, with p-values reported to substantiate the efficacy of EMFL.

Handling Neutral Sentiments: Given the scarcity of neutral sentiment utterances across all datasets, neutral data points were excluded from the experiments. This decision focuses the evaluation on binary sentiment classification (positive and negative), which is more prevalent and statistically robust.

Baseline Verification: To ensure the reliability of our textual feature extraction, we compared the CNN's performance against several recently published sentiment analysis tools trained on large-scale movie reviews. The testing accuracy of these tools ranged from 0.68 to 0.70, which was slightly lower than the CNN's validation accuracy of 0.724, thereby validating the robustness of our baseline model.

4.5. Experimental Results

4.5.1. Within-Dataset Evaluation

Table 2 presents the classification accuracy for both CNN and EMFL-CNN models evaluated within the MOSI dataset, specifically on the subset of 31 individuals excluded from the training and validation phases. The results unequivocally demonstrate that the integration of EMFL consistently enhances the model's performance across most modalities. Notably, the EMFL-CNN achieves superior accuracy in verbal, acoustic, and visual modalities individually, as well as in most bimodal combinations. This improvement underscores EMFL's effectiveness in mitigating the influence of confounding identity-related factors, thereby enabling the model to focus more intently on sentiment-relevant features.

 doi:10.20944/preprints202409.1887.v

Table 2. Classification accuracy within the MOSI dataset for CNN and EMFL-CNN models across various modalities. The EMFL-CNN consistently outperforms the baseline CNN, demonstrating enhanced generalizability within the same dataset.

		CNN	EMFL-CNN
Unimodal	Verbal	0.678	**0.732**
	Acoustic	0.588	**0.618**
	Visual	0.572	**0.636**
Bimodal	Verbal+Acoustic	0.687	**0.725**
	Verbal+Visual	0.706	**0.73**
	Acoustic+Visual	**0.661**	0.621
All Modalities		0.715	**0.73**

4.5.2. Cross-Dataset Evaluation

Table 3 delineates the performance of CNN and EMFL-CNN models on the YouTube and MOUD test datasets. The findings reveal that the CNN baseline occasionally underperforms, with some accuracies falling below chance levels. This underperformance substantiates the presence of generalization challenges when models trained on one dataset are applied to others with differing characteristics.

Table 3. Classification accuracy across different datasets (YouTube and MOUD) for CNN and EMFL-CNN models across various modalities. The EMFL-CNN exhibits significant performance gains, highlighting its enhanced generalizability across diverse data distributions.

	YouTube		MOUD	
	CNN	**EMFL-CNN**	**CNN**	**EMFL-CNN**
Verbal	0.605	**0.657**	0.522	**0.569**
Acoustic	0.441	**0.564**	0.455	**0.549**
Visual	0.492	**0.549**	**0.555**	0.548
Verbal+Acoustic	0.642	**0.652**	0.515	**0.574**
Verbal+Visual	0.642	**0.667**	0.542	**0.574**
Acoustic+Visual	0.452	**0.559**	0.533	**0.554**
All Modalities	0.611	**0.667**	0.531	**0.574**

The EMFL-CNN model consistently outperforms the CNN baseline across almost all modalities and dataset combinations. Specifically, in the YouTube dataset, EMFL-CNN improves verbal, acoustic, and visual modality accuracies from 0.605 to 0.657, 0.441 to 0.564, and 0.492 to 0.549, respectively. Similarly, in the MOUD dataset, EMFL-CNN enhances accuracies from 0.522 to 0.569 in the verbal modality and from 0.455 to 0.549 in the acoustic modality, while achieving competitive performance in the visual modality.

Modality-Specific Insights

Textual modality consistently exhibits the highest classification accuracy across both within-dataset and cross-dataset evaluations. This superiority can be attributed to two primary factors:

1. **Semantic Richness**: Sentiment is inherently more nuanced and accurately captured through textual expressions, which provide explicit cues compared to the more ambiguous visual or acoustic signals.
2. **Language Independence from Identity**: Textual information is less likely to be influenced by an individual's identity, reducing the risk of confounding factors impacting sentiment prediction.

However, despite the robustness of textual features, our results indicate that even textual modalities can be susceptible to confounding by individual-specific language preferences and stylistic variations, particularly in spoken language contexts.

The disparate nature of the YouTube and MOUD datasets, despite originating from the same web platform, introduces variations in recording quality and post-processing techniques. These

 doi:10.20944/preprints202409.1887.v

discrepancies manifest as differing data distributions, adversely affecting the CNN's performance in acoustic and visual modalities. For instance, in the acoustic modality, CNN accuracy on YouTube drops to 0.441 from within-dataset training, while EMFL-CNN achieves a substantial improvement to 0.564. Similar trends are observed in the visual modality, where EMFL-CNN maintains competitive performance despite inherent dataset variations.

In multimodal settings, combining features from multiple modalities generally enhances sentiment prediction accuracy. Our experiments confirm that EMFL-CNN consistently improves multimodal fusion results across both within-dataset and cross-dataset evaluations. For example, in the MOSI dataset, integrating verbal and visual modalities yields an accuracy of 0.73 with EMFL-CNN compared to 0.706 with CNN. This enhancement underscores EMFL's capability to effectively disentangle and leverage complementary information from multiple modalities, further boosting model robustness and accuracy.

To validate the significance of the observed performance improvements, we conducted permutation tests comparing the CNN and EMFL-CNN models. The null hypothesis posits no improvement with EMFL-CNN. The resultant p-values—0.037 for MOSI, 0.0003 for YouTube, and 0.0023 for MOUD—reject the null hypothesis, thereby confirming that the performance gains achieved by EMFL-CNN are statistically significant.

While EMFL-CNN generally outperforms the CNN baseline, there are a few exceptions. Specifically, in the MOUD dataset's visual modality and the MOSI dataset's acoustic and visual bimodal combinations, CNN occasionally matches or slightly outperforms EMFL-CNN. These instances may arise from dataset-specific nuances or the inherent limitations of EMFL in certain multimodal configurations. Further investigation is warranted to understand and address these anomalies, potentially through model refinement or additional confounder mitigation strategies.

The experimental results robustly demonstrate that the EMFL framework significantly enhances the generalizability and robustness of discriminative neural classifiers across diverse datasets and modalities. By effectively mitigating the influence of identity-related confounders, EMFL-CNN models exhibit superior performance, particularly in challenging cross-dataset scenarios where data distributions vary substantially. These findings validate the efficacy of EMFL in advancing the state-of-the-art in multimodal sentiment analysis.

While the current study focuses on identity-related confounders, future research could explore the extension of the EMFL framework to account for multiple types of confounders, such as contextual or environmental factors. Additionally, integrating advanced feature extraction techniques and exploring deeper neural architectures could further bolster the framework's performance. Expanding evaluations to include more diverse and larger-scale datasets would also provide a more comprehensive assessment of EMFL's generalizability across various real-world applications.

5. Conclusions and Future Work

5.1. Conclusion

In the realm of automatic multimodal sentiment analysis, the scarcity of high-quality datasets poses a significant challenge. Typically, the datasets available for training machine learning models consist of only a few thousand samples. This limitation inherently restricts the models' ability to generalize effectively across diverse and unseen data, primarily due to the presence of confounding factors that can bias the sentiment predictions. These confounders, often stemming from identity-related features such as individual speaking styles or unique visual cues, can detract from the model's focus on sentiment-associated features, thereby diminishing overall performance and reliability.

To address this critical issue, we introduced the Enhanced Multimodal Feature Learning (EMFL) framework, a novel approach designed to mitigate the adverse effects of confounding factors on sentiment analysis models. EMFL operates by identifying and suppressing identity-related confounding dimensions within the feature space, thereby allowing the model to concentrate more accurately on features that are genuinely indicative of sentiment. This selective enhancement ensures that the

 doi:10.20944/preprints202409.1887.v

model's predictive capabilities are anchored in sentiment-relevant information, rather than being inadvertently influenced by irrelevant identity-specific attributes.

Our extensive experimental evaluations across three distinct datasets—MOSI, YouTube, and MOUD—demonstrate the efficacy of the EMFL framework in significantly improving the generalizability of state-of-the-art multimodal sentiment analysis models. The results indicate substantial gains in prediction accuracy across all three modalities: verbal, acoustic, and visual. Furthermore, the integration of EMFL into existing models not only enhances unimodal performance but also fortifies multimodal fusion strategies, leading to more robust and reliable sentiment predictions.

A key highlight of our findings is EMFL's ability to maintain high prediction accuracy even when models are tested across different datasets. This cross-dataset robustness underscores EMFL's effectiveness in overcoming dataset-specific biases and variations, thereby ensuring that the models remain resilient and perform consistently well in diverse real-world scenarios. Such generalizability is crucial for the deployment of sentiment analysis systems in dynamic environments where data distributions can vary widely.

In summary, the Enhanced Multimodal Feature Learning (EMFL) framework offers a significant advancement in the field of multimodal sentiment analysis by effectively addressing the limitations imposed by limited and confounded datasets. By fostering a more focused and discriminative feature learning process, EMFL not only enhances the accuracy of sentiment predictions but also ensures that these models are better equipped to generalize across varied and unseen data landscapes.

5.2. Future Work

While the EMFL framework has demonstrated considerable improvements in mitigating confounding factors and enhancing model generalizability, there remain several avenues for future research to further refine and extend its capabilities.

Firstly, exploring the application of EMFL to larger and more diverse datasets could provide deeper insights into its scalability and effectiveness in even more complex sentiment analysis tasks. Additionally, integrating EMFL with more advanced neural architectures, such as Transformer-based models, may unlock further performance gains and adaptability across different modalities.

Secondly, extending the framework to handle multiple types of confounding factors beyond identity-related features could broaden its applicability. For instance, environmental factors or contextual elements that influence sentiment expression could be incorporated into the model, enabling a more comprehensive disentanglement of relevant and irrelevant features.

Thirdly, investigating semi-supervised or unsupervised variants of EMFL could alleviate the dependency on labeled data, making the framework more versatile in scenarios where annotated datasets are scarce or costly to obtain. This approach could leverage unlabeled data to further enhance feature learning and model robustness.

Lastly, applying the EMFL framework to other domains beyond sentiment analysis, such as emotion recognition, intent detection, or human-computer interaction, could demonstrate its versatility and potential for widespread impact across various fields. Each of these directions presents an opportunity to build upon the foundational strengths of EMFL, driving forward the capabilities of multimodal machine learning models.

By addressing these future research directions, the EMFL framework can continue to evolve, contributing to the development of more sophisticated, reliable, and generalizable multimodal sentiment analysis systems.

References

1. Louis-Philippe Morency, Rada Mihalcea, and Payal Doshi, "Towards multimodal sentiment analysis: Harvesting opinions from the web," in *Proceedings of the 13th international conference on multimodal interfaces*. ACM, 2011.

2. Bo Pang and Lillian Lee, "Opinion mining and sentiment analysis," *Foundations and trends in information retrieval*, 2008.
3. Akshi Kumar and Mary Sebastian Teeja, "Sentiment analysis: A perspective on its past, present and future," *International Journal of Intelligent Systems and Applications*, 2012.
4. Martin Wollmer, Felix Weninger, Timo Knaup, Bjorn Schuller, Congkai Sun, Kenji Sagae, and Louis-Philippe Morency, "Youtube movie reviews: Sentiment analysis in an audio-visual context," *Intelligent Systems, IEEE*, 2013.
5. Verónica Pérez Rosas, Rada Mihalcea, and Louis-Philippe Morency, "Multimodal sentiment analysis of spanish online videos," *IEEE Intelligent Systems*, , no. 3, 2013.
6. Amir Zadeh, "Micro-opinion sentiment intensity analysis and summarization in online videos," in *ICMI*. ACM, 2015.
7. Robert M Ewers and Raphael K Didham, "Confounding factors in the detection of species responses to habitat fragmentation," *Biological Reviews*, 2006.
8. Haohan Wang and Jingkang Yang, "Multiple confounders correction with regularized linear mixed effect models, with application in biological processes," in *BIBM*. IEEE, 2016.
9. Lingxiang Wu, Jinqiao Wang, Guibo Zhu, Min Xu, and Hanqing Lu, "Person re-identification via rich color-gradient feature," in *ICME*. IEEE, 2016.
10. Xiaoke Zhu, Xiao-Yuan Jing, Fei Wu, Weishi Zheng, Ruimin Hu, Chunxia Xiao, and Chao Liang, "Distance learning by treating negative samples differently and exploiting impostors with symmetric triplet constraint for person re-identification," in *ICME*. IEEE, 2016.
11. Antonio Tejero-de Pablos, Yuta Nakashima, Tomokazu Sato, and Naokazu Yokoya, "Human action recognition-based video summarization for rgb-d personal sports video," in *ICME*. IEEE, 2016.
12. Ying Zhao, Huijun Di, Jian Zhang, Yao Lu, and Feng Lv, "Recognizing human actions from low-resolution videos by region-based mixture models," in *ICME*. IEEE, 2016.
13. Zhongjun Wu and Weihong Deng, "One-shot deep neural network for pose and illumination normalization face recognition," in *ICME*. IEEE, 2016.
14. Binghui Chen and Weihong Deng, "Weakly-supervised deep self-learning for face recognition," in *ICME*. IEEE, 2016.
15. Erik Cambria, Daniel Olsher, and Dheeraj Rajagopal, "Senticnet 3: a common and common-sense knowledge base for cognition-driven sentiment analysis," in *AAAI*. AAAI Press, 2014.
16. Theresa Wilson, Janyce Wiebe, and Paul Hoffmann, "Recognizing contextual polarity in phrase-level sentiment analysis," in *EMNLP*. Association for Computational Linguistics, 2005.
17. Ellen Riloff and Janyce Wiebe, "Learning extraction patterns for subjective expressions," in *EMNLP*. Association for Computational Linguistics, 2003.
18. Lakshmish Kaushik, Abhijeet Sangwan, and John HL Hansen, "Sentiment extraction from natural audio streams," in *ICASSP*. IEEE, 2013.
19. Boya Wu, Jia Jia, Tao He, Juan Du, Xiaoyuan Yi, and Yishuang Ning, "Inferring users'emotions for human-mobile voice dialogue applications," .
20. Paul Ekman and Wallace V Friesen, "Facial action coding system," 1977.
21. Ming Sun, Jufeng Yang, Kai Wang, and Hui Shen, "Discovering affective regions in deep convolutional neural networks for visual sentiment prediction," in *ICME*. IEEE, 2016.
22. Verónica Pérez-Rosas, Rada Mihalcea, and Louis-Philippe Morency, "Utterance-level multimodal sentiment analysis.," in *ACL*, 2013.
23. Luca Casaburi, Francesco Colace, Massimo De Santo, and Luca Greco, ""magic mirror in my hand, what is the sentiment in the lens?": An action unit based approach for mining sentiments from multimedia contents," *Journal of Visual Languages & Computing*.
24. Soujanya Poria, Erik Cambria, Newton Howard, Guang-Bin Huang, and Amir Hussain, "Fusing audio, visual and textual clues for sentiment analysis from multimodal content," *Neurocomputing*.
25. Soujanya Poria, Erik Cambria, and Alexander Gelbukh, "Deep convolutional neural network textual features and multiple kernel learning for utterance-level multimodal sentiment analysis," in *EMNLP*, 2015, pp. 2539–2544.
26. Diederik P Kingma and Max Welling, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013.

 doi:10.20944/preprints202409.1887.v

27. Haohan Wang and Bhiksha Raj, "On the origin of deep learning," *arXiv preprint arXiv:1702.07800*, 2017.
28. Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean, "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781*, 2013.
29. Florian Eyben, Martin Wöllmer, and Björn Schuller, "Opensmile: the munich versatile and fast open-source audio feature extractor," in *International conference on Multimedia*. ACM, 2010.
30. Tadas Baltrušaitis, Peter Robinson, and Louis-Philippe Morency, "3d constrained local model for rigid and non-rigid facial tracking," in *CVPR*. IEEE, 2012, pp. 2610–2617.
31. Anson Bastos, Abhishek Nadgeri, Kuldeep Singh, Isaiah Onando Mulang, Saeedeh Shekarpour, Johannes Hoffart, and Manohar Kaul. 2021. RECON: Relation Extraction using Knowledge Graph Context in a Graph Neural Network. In *Proceedings of the Web Conference 2021*. 1673–1685.
32. Philipp Christmann, Rishiraj Saha Roy, Abdalghani Abujabal, Jyotsna Singh, and Gerhard Weikum. 2019. Look before You Hop: Conversational Question Answering over Knowledge Graphs Using Judicious Context Expansion. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management CIKM*. 729–738.
33. 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.
34. 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).
35. 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.
36. 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.
37. 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)*.
38. 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.
39. Ilya Loshchilov and Frank Hutter. 2019. Decoupled Weight Decay Regularization. In *International Conference on Learning Representations*.
40. 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).
41. Matthew J Smith. Getting value from artificial intelligence in agriculture. *Animal Production Science*, 2018.
42. 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.
43. 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.
44. 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.
45. 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.

 doi:10.20944/preprints202409.1887.v

46. 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.
47. 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.
48. Ling Zhuang, Hao Fei, and Po Hu. Knowledge-enhanced event relation extraction via event ontology prompt. *Inf. Fusion*, 100:101919, 2023.
49. 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.
50. 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.
51. 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.
52. 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.
53. 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.
54. 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.
55. 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.
56. 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.
57. Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*, 2016.
58. 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.
59. 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. 10.18653/v1/N19-1423. URL https://aclanthology.org/N19-1423.
60. Shengqiong Wu, Hao Fei, Leigang Qu, Wei Ji, and Tat-Seng Chua. Next-gpt: Any-to-any multimodal llm. *CoRR*, abs/2309.05519, 2023.
61. 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.
62. 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.
63. Naman Jain, Pranjali Jain, Pratik Kayal, Jayakrishna Sahit, Soham Pachpande, Jayesh Choudhari, et al. Agribot: agriculture-specific question answer system. *IndiaRxiv*, 2019.
64. 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.
65. 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.

66. 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.
67. 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.
68. 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.
69. 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.
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. 2024.
71. Sanjeev Arora, Yingyu Liang, and Tengyu Ma. A simple but tough-to-beat baseline for sentence embeddings. In *ICLR*, 2017.
72. 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.
73. 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.
74. 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.
75. 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.
76. 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.
77. 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.
78. 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.
79. 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.
80. 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.
81. 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.
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.

 doi:10.20944/preprints202409.1887.v

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. 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.