Article

Not peer-reviewed version

Multigranular Unified Synthesis Encoder for Fine-grained Multimodal Emotion Understanding

Colton Ray , Wyne Nasir , Savannah Grace *

Posted Date: 16 May 2025

doi: 10.20944/preprints202505.1230.v1

Keywords: multigranular fusion; multimodal emotion recognition; unified synthesis encoder; fine-grained cross-modal interaction; pre-trained representations

Article

Multigranular Unified Synthesis Encoder for Fine-Grained Multimodal Emotion Understanding

Colton Ray, Wyne Nasir and Savannah Grace *

Tufts University
* Correspondence: savannahg@tufts.edu

Abstract: Accurate emotion understanding from multimodal signals has become a pivotal research area, especially given its relevance in enhancing human-computer interaction systems. However, the inherent complexity of emotional expression across modalities, coupled with the scarcity of high-quality annotated data, poses significant barriers to progress. In this work, we present MUSE, a novel multigranular unified synthesis encoder framework, designed to seamlessly integrate fine-grained representations and global pre-trained embeddings for superior emotion recognition. In contrast to prior studies which narrowly emphasize either modality-level pretraining or local feature alignment, our method orchestrates both perspectives synergistically. Drawing inspiration from advances in text-to-speech synthesis, MUSE employs a multilevel Transformer-based module that explicitly models cross-modal associations among phonemes, words, and utterances. Furthermore, it leverages self-supervised learning backbones to exploit large-scale unlabeled corpora efficiently. Our extensive evaluations on the widely adopted IEMOCAP benchmark reveal that MUSE consistently surpasses existing approaches, establishing new state-of-the-art performances. Additionally, we demonstrate that our multigranular fusion strategy yields substantial gains over conventional fusion schemes.

Keywords: multigranular fusion; multimodal emotion recognition; unified synthesis encoder; fine-grained cross-modal interaction; pre-trained representations

1. Introduction

Emotion recognition from speech remains a core challenge within the broader context of affective computing, particularly for enabling more nuanced and empathetic human-computer dialogue systems. The task, commonly known as Speech Emotion Recognition (SER), focuses on inferring the speaker's affective states such as joy, anger, or sorrow through vocal expressions [1]. Despite its potential, SER research faces two fundamental bottlenecks. Firstly, obtaining large-scale labeled emotion datasets is notoriously difficult due to the subjectivity of emotional perception and the demanding nature of multi-annotator labeling processes [2]. Secondly, emotional expression naturally spans multiple modalities, and these modalities exhibit subtle, fine-grained interactions that are challenging to capture and model effectively [3].

A prevalent strategy to address the data limitation is to employ transfer learning paradigms, particularly those driven by self-supervised learning (SSL). SSL has emerged as a game-changing approach, achieving impressive results in NLP [4–6] and speech domains [7–9]. Within the realm of emotion recognition, pioneering efforts by Acheampong et al. [10] and Pepino et al. [11] have independently leveraged text and speech pre-trained models, yielding promising results. Yet, these endeavors restrict themselves to unimodal processing.

Parallel efforts have sought to enhance multimodal integration by exploring varied fusion strategies. Siriwardhana et al. [12] and Zhao et al. [2] compared early and late fusion schemes combining BERT [4] and Wav2Vec [7] for emotion classification, revealing that late fusion generally achieves better outcomes. However, these methods relied solely on aggregated sentence-level embeddings, neglecting the finer-grained elements such as phonemes or word-level cues. Conversely, models devoid of pre-trained backbones utilized diverse architectures [3,13,14] to capture modality interactions.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 16 May 2025 doi:10.20944/preprints202505.1230.v1

Notably, Yoon et al. [13] introduced an RNN-based approach to model speech and text separately before concatenating them for classification, albeit at the utterance level. Xu et al. [14] proposed an alignment strategy using LSTM to jointly process text and speech fragments. However, such sequential modeling restricted intramodal interaction exploration. Li et al. [3] developed a fine-grained method using temporal alignment with mean-max pooling and cross-modal mechanisms but introduced extra overhead in alignment prediction, complicating its deployment in practical systems.

While SSL frameworks empower the derivation of robust embeddings from massive unlabeled datasets [12], these representations tend to encode holistic sentence-level semantics rather than fine-grained elements like specific words or phonetic nuances. Addressing this gap necessitates innovative approaches that incorporate granular cross-modal interactions without imposing additional labeling burdens on annotators. Transformer TTS [15] exemplifies such fine-grained modeling in text-to-speech synthesis, employing phoneme and mel spectrogram sequences as inputs to generate speech outputs. Drawing parallels from this domain, we hypothesize that analogous architectures can be repurposed for SER, enabling refined modeling of audio-text interactions at the phoneme level, circumventing explicit alignment steps. Nevertheless, relying solely on phoneme-level inputs falls short in capturing the holistic meaning provided by word-level semantics.

Motivated by these challenges, we propose MUSE, a holistic multigranular framework designed to amalgamate utterance-level pre-trained embeddings with fine-grained representations. At its core, MUSE leverages a specialized multilevel Transformer module to instill cross-modal alignment among voice fragments, words, and phonemes, while integrating phoneme embeddings with their corresponding word embeddings through sophisticated strategies. Additionally, a vanilla Transformer encoder [16] is incorporated to reinforce the sequential multimodal representation aggregation. For global semantics, BERT [4] serves as our chosen pre-trained model backbone, facilitating multigranular fusion. Our rigorous experimental validations on the IEMOCAP dataset [17] demonstrate that MUSE consistently surpasses contemporary state-of-the-art methods, with its multigranular design delivering tangible performance uplift.

To summarize, our principal contributions are outlined as follows:

- We introduce a novel multilevel Transformer-based encoder, MUSE, capable of modeling intricate cross-modal interactions at the phoneme, word, and utterance levels (Section 3.1), fostering enriched multimodal feature representations.
- We devise an efficient yet effective multigranular fusion paradigm that seamlessly bridges fine-grained and pre-trained utterance-level representations, enhancing emotion understanding fidelity (Section 3.2).
- We comprehensively evaluate MUSE on the IEMOCAP benchmark, wherein it achieves superior results over existing approaches, validating the efficacy of our proposed multigranular fusion strategy (Section 4).

2. Related Work

The evolution of speech emotion recognition (SER) has witnessed a significant paradigm shift from traditional handcrafted feature-based methods to contemporary deep learning-driven approaches. Early research predominantly relied on classical machine learning algorithms such as the Hidden Markov Model (HMM) [18] and the Gaussian Mixture Model (GMM) [19], where carefully designed low-level acoustic descriptors and statistical high-level features formed the foundation for emotion classification tasks. These conventional methods, while providing initial insights into the field, were inherently limited by their dependence on shallow representations and lacked the capacity to capture complex, hierarchical patterns present in emotional speech data.

To address these constraints, the advent of deep neural networks (DNNs) introduced transformative capabilities in SER. D. Bertero et al. [20] pioneered the utilization of convolutional neural networks (CNNs) to automatically extract abstract feature representations directly from raw spectrogram inputs, facilitating end-to-end learning without the necessity of manual feature engineering. Similarly, A.

Satt et al. [21] proposed a hybrid framework combining CNNs and long short-term memory (LSTM) networks, effectively capturing both spatial and temporal emotional cues. This approach enabled the model to learn hierarchical feature structures while simultaneously modeling contextual dependencies inherent in speech sequences.

With the growing realization that human emotional expression is inherently multimodal, recent research has expanded towards integrating both auditory and textual modalities to enrich emotion recognition systems. For example, S. Yoon et al. [13] developed a multimodal SER model employing recurrent neural networks (RNNs) to encode both audio and text streams independently. Their approach involved utilizing the final hidden state of one modality's encoder as a query in an attention mechanism applied to the other modality's encoder outputs, thereby attempting to bridge the two modalities during fusion. Nevertheless, their method did not fully exploit bidirectional inter-modality interactions, and the attention mechanism primarily focused on global representations rather than exploring fine-grained token-level or phoneme-level interactions.

In another line of work, H. Xu et al. [14] proposed a model leveraging LSTM-based encoders alongside an attention mechanism to align and integrate audio and text modalities. Although this method introduced a more explicit alignment between modalities, it remained restricted by the sequential nature of LSTM, which processes information linearly and may suffer from information loss over long sequences. Moreover, their approach did not sufficiently explore intramodal dynamics, potentially neglecting important cues within the individual modalities themselves.

Further pushing the boundaries, H. Li et al. [3] introduced a fine-grained multimodal emotion recognition framework by incorporating a temporal alignment mean-max pooling strategy. This method facilitated more granular interactions between audio and text, allowing the model to focus on localized temporal segments where emotional cues are most prominent. In addition, they integrated a cross-modality interaction module to enhance feature fusion. However, their framework required explicitly aligned audio and text inputs, which imposes significant preprocessing overhead and limits scalability in real-world scenarios where such alignment information might not be readily available or reliable.

Despite these advancements, current approaches still face limitations in capturing the subtle nuances of multimodal emotional expression. The primary shortcoming lies in the insufficient modeling of cross-modal and intra-modal interactions at different granularity levels, including utterance, word, and phoneme levels. This motivates the need for more holistic frameworks that can flexibly incorporate fine-grained interactions while leveraging the robustness of pre-trained representations.

Our proposed MUSE framework draws inspiration from these prior works yet overcomes their limitations by introducing a unified multigranular synthesis encoder that explicitly models cross-modal interactions at multiple granularity levels. Through the integration of a hierarchical Transformer-based encoder and a multigranular fusion module, MUSE allows for seamless blending of pre-trained utterance-level embeddings with fine-grained audio-text representations without requiring additional alignment annotations. Furthermore, by employing advanced attention mechanisms that dynamically capture both inter-modality and intra-modality dependencies, MUSE offers a more comprehensive approach for multimodal emotion understanding.

In contrast to earlier models that relied solely on either early or late fusion strategies, our method harmonizes both strategies within a unified learning framework, offering a more adaptive and fine-grained fusion scheme. Moreover, by leveraging recent advances in self-supervised learning (SSL) and Transformer architectures, MUSE capitalizes on large-scale unannotated data to pretrain its encoders, thereby mitigating the reliance on scarce labeled datasets. This allows the framework to generalize more effectively across diverse emotional contexts and speaker variabilities.

In summary, while the field of multimodal emotion recognition has made remarkable progress over the years, our review of the literature highlights persistent gaps in modeling capabilities, particularly in handling fine-grained interactions and reducing the dependency on explicit alignment information. MUSE is designed to address these gaps by synergizing fine-grained feature modeling with global semantic representations, paving the way for more robust and scalable SER systems.

3. Framework

In this section, we provide a comprehensive introduction to our proposed **MUSE** (Multigranular Unified Synthesis Encoder) framework. First, we detail the novel multilevel Transformer-based module designed for fine-grained multimodal emotion understanding. Subsequently, we elaborate on the overall multi-granularity fusion strategy, which unifies our proposed module with powerful pre-trained language representations.

3.1. Multilevel Transformer Encoder for Fine-grained Interaction

3.1.1. Model Design and Workflow

Inspired by the architecture of Transformer TTS [15], which itself is based on the foundations of Tacotron2 [22] and the Transformer model [16], we present a novel multilevel Transformer module within **MUSE**. This module is designed to systematically model the intricate cross-modal dependencies between phonemes, words, and acoustic representations.

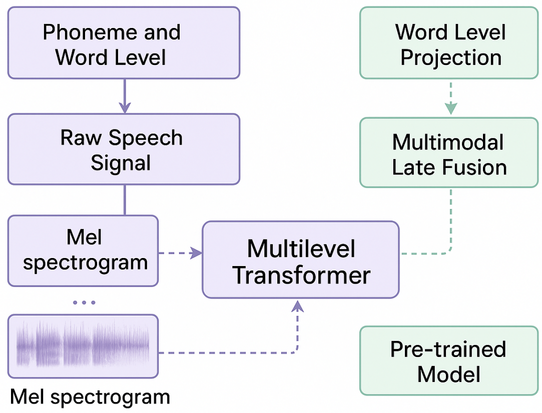

Figure 1. Overview of Multilevel Transformer.

Initially, the text input undergoes decomposition into both phoneme and word sequences, ensuring the extraction of both micro and macro linguistic cues. These are subsequently passed through a highway network [23] to enhance gradient flow and prevent information bottlenecks. Following this, an encoder pre-net composed of three convolutional layers and a projection layer processes the textual embeddings into a format suitable for sequential encoding. Simultaneously, the mel spectrogram, capturing the speech signal, is processed through a two-layer fully connected network to generate a condensed representation.

To enable rich cross-modal interaction, the processed text and audio features are fed into a *Cross-Modal Attention Interaction Module*, followed by a *Deep Fusion Module*, both of which leverage the vanilla Transformer architecture. Finally, a classification head operates on the aggregated representation derived from the dummy mel input vector, predicting the probability distribution over emotion categories.

3.1.2. Sequential Interaction Modeling and Cross-modal Encoding

During both training and inference phases, the system consumes the complete mel spectrogram and textual information, ensuring maximal context utilization. A dummy mel vector m_{dummy} is prepended to the mel spectrogram sequence, yielding $m = (m_{dummy}, m_1, m_2, ..., m_{T'})$. This enables the model to derive a global representation from the sequence, akin to the [CLS] token in BERT-based models.

The sequence-to-sequence nature of TTS tasks can be formalized as:

$$f(o_t|x_1, ..., x_T) = f(o_t|o_{<t}, x) \tag{1}$$

However, in our scenario, the target is not mel spectrogram prediction but rather emotion classification, formalized as:

$$p = g(x, m) \tag{2}$$

where $g(x, m)$ denotes the classification function leveraging both modalities.

3.1.3. Hierarchical Phoneme and Word Embedding Strategy

Recognizing the nuanced contribution of phonemes in emotion cues, we apply a convolutional neural network (CNN) to extract fixed-dimensional phoneme embeddings for each word following [24,25]. Specifically:

$$e^i_{phoneme} = \text{MaxPool}(\text{CNN}(\text{PhonemeSeq}_i)) \tag{3}$$

In parallel, we incorporate word-level semantic embeddings derived from GloVe [26]:

$$e^i_{word} = \text{GloVe}(w_i) \tag{4}$$

3.1.4. Fusion of Phoneme and Word Embeddings

We explore two approaches to integrate phoneme and word embeddings:

1) Concatenation Strategy: A naive yet effective method involves direct concatenation:

$$u_i = [e^i_{phoneme}; e^i_{word}] \tag{5}$$

2) Highway Network-based Fusion: Inspired by [23], the fused vector is passed through a two-layer highway network:

$$Z(u) = H(u) \cdot T(u) + u \cdot (1 - T(u)) \tag{6}$$

where $H(u) = \text{ReLU}(W_h u + b_h)$ is the transformation function, and $T(u) = \sigma(W_t u + b_t)$ is the transform gate controlling information flow.

3.1.5. Cross-modal Attention and Deep Fusion Modules

Each module adopts a vanilla Transformer block comprising multi-head self-attention, position-wise feed-forward layers, and residual connections [16]. The Cross-Modal Attention module uses encoder-decoder attention to condition the mel features on the textual embeddings, defined as:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{7}$$

where Q are the query vectors from mel representations, and K, V are key and value vectors from text encodings.

The Deep Fusion Module aggregates these multimodal features through additional self-attention layers, enhancing both intra and inter-modal interactions.

3.1.6. Objective Functions and Multi-task Considerations

While inspired by multitask learning strategies from [28], we empirically observed no gains from joint TTS and SER loss optimization in our setting. Thus, we optimize only the cross-entropy loss for emotion classification:

$$L_{SER} = -\sum_{i=1}^{N}\sum_{k=1}^{K} y_{i,k} \log p_{i,k} \tag{8}$$

 doi:10.20944/preprints202505.1230.v1

We further regularize our model using auxiliary attention alignment loss:

$$L_{align} = \frac{1}{T \times T'} \sum_{t=1}^{T} \sum_{t'=1}^{T'} \mathrm{KL}(A_{t,t'} || \hat{A}_{t,t'}) \tag{9}$$

where $A_{t,t'}$ is the predicted attention matrix and $\hat{A}_{t,t'}$ is the ideal alignment.

3.2. Multigranular Fusion Network

3.2.1. Global and Local Representation Integration

Our multigranular framework leverages both global sentence-level embeddings from pre-trained BERT and the fine-grained multimodal representation from the multilevel Transformer encoder. BERT's [CLS] token is used as the global semantic summarization of the text sequence:

$$e_{BERT} = \mathrm{BERT}_{CLS}(x) \tag{10}$$

Concurrently, the first vector from the Deep Fusion Module output represents the fine-grained cross-modal embedding:

$$e_{MUSE} = \mathrm{DFM}_{[1]}(m, x) \tag{11}$$

3.2.2. Fusion Strategy and Classification Head

Adopting a late fusion strategy inspired by [12,29], we perform:

$$e_{concat} = [\mathrm{Proj}(e_{BERT}); \mathrm{Proj}(e_{MUSE})] \tag{12}$$

This concatenated representation is passed through a feed-forward classifier:

$$p = \mathrm{softmax}(W_c e_{concat} + b_c) \tag{13}$$

where W_c and b_c are learnable parameters, and p is the predicted emotion category distribution.

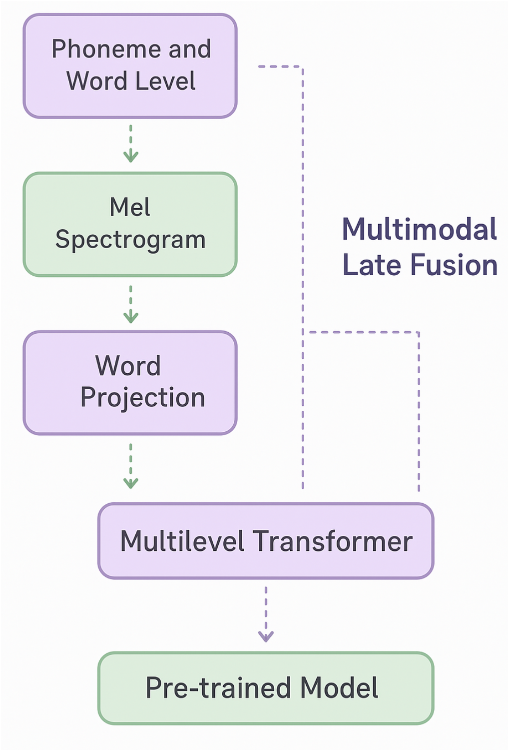

Figure 2. Illustration of Multigranular Fusion Network.

 doi:10.20944/preprints202505.1230.v1

3.2.3. Enhanced Objective Function with Regularization

In addition to L_{SER} and L_{align}, we introduce an auxiliary embedding alignment loss ensuring the coherence of e_{BERT} and e_{MUSE}:

$$L_{emb\_align} = \|e_{BERT} - e_{MUSE}\|_2^2 \tag{14}$$

Our final loss is:

$$L_{total} = L_{SER} + \lambda_1 L_{align} + \lambda_2 L_{emb\_align} \tag{15}$$

where λ_1 and λ_2 are balancing hyperparameters.

4. Experiments

In this section, we conduct extensive experimental evaluations to validate the effectiveness and robustness of our proposed **MUSE** framework. We systematically describe the dataset utilized, elaborate the implementation configurations, and present comprehensive performance comparisons with state-of-the-art methods. Additionally, we perform an ablation study to investigate the contribution of different modules within the MUSE architecture. All experimental results are analyzed both quantitatively and qualitatively to offer in-depth insights.

4.1. Dataset and Experimental Settings

To comprehensively assess the performance of our models, we adopt the widely-used Interactive Emotional Dyadic Motion Capture (IEMOCAP) dataset [17], a benchmark dataset extensively applied in multimodal emotion recognition research. This dataset comprises approximately 12 hours of audiovisual recordings, including speech, video, and textual transcripts. In this study, we exclusively utilize the speech audio and corresponding text transcripts.

Following the prevalent experimental protocol proposed by prior works such as [13], we consider four core emotion categories: *angry* (1103 samples), *sad* (1084 samples), *neutral* (1708 samples), and *happy* (merged with *excited*, totaling 1636 samples), resulting in an overall dataset containing 5531 utterances. For experimental consistency, we apply 5-fold speaker-independent cross-validation with splits configured as 60% for training, 20% for development, and 20% for testing. Each experiment is conducted thrice with distinct random seeds, and the averaged performance is reported to mitigate randomness bias.

4.2. Implementation Details and Training Protocol

All models are implemented using the PyTorch deep learning framework. For acoustic feature extraction, 128-dimensional filterbank features are computed from raw speech waveforms using a window size of 25ms and a hop size of 12ms. For textual modality, we utilize 300-dimensional pre-trained GloVe embeddings [26] to represent words. All Transformer modules within our MUSE framework are configured with a hidden size of 128 and utilize multi-head self-attention with 4 heads.

The models are optimized using the Adam optimizer [30] with an initial learning rate set to $1e^{-5}$ and a mini-batch size of 4. We employ early stopping based on the best validation Weighted Accuracy (WA) over 10 epochs. Final model evaluations are performed using both WA and Unweighted Accuracy (UA) metrics on the held-out test set to provide a comprehensive understanding of performance across class distributions.

4.3. Evaluation of MUSE's Multilevel Transformer Component

We perform an extensive evaluation of our multilevel transformer module embedded within MUSE on the IEMOCAP dataset. Table 1 reports the comparative results. It is evident that the proposed multilevel transformer model significantly outperforms all previous baselines in both WA and UA. Notably, our approach surpasses the current best method [3], achieving a WA of 0.735 and UA of

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 16 May 2025 doi:10.20944/preprints202505.1230.v1

0.747, demonstrating the effectiveness of fine-grained cross-modal modeling without relying on costly alignment annotations.

The ablation analysis further reveals that word embeddings contribute more substantially to the recognition task compared to phoneme-only representations, suggesting the importance of semantic context. Furthermore, integrating both word and phoneme embeddings through a highway network achieves superior performance, validating the design choice of using highway networks to enable efficient information fusion. Additionally, the absence of the Deep Fusion module leads to observable performance degradation, emphasizing its crucial role in capturing deep multimodal dependencies.

Table 1. Performance comparison of the MUSE multilevel transformer model against state-of-the-art baselines.

Methods	**WA**	**UA**
S. Yoon et al. [13]	0.682 ± 0.012	0.688 ± 0.014
H. Xu et al. [14]	0.685 ± 0.007	0.691 ± 0.008
H. Li et al. [3]	0.716 ± 0.004	0.725 ± 0.005
MUSE Multilevel Transformer	$\mathbf{0.735 \pm 0.004}$	$\mathbf{0.747 \pm 0.003}$
Ablation Study	**WA**	**UA**
Phoneme only	0.680 ± 0.003	0.695 ± 0.005
Word only	0.715 ± 0.002	0.726 ± 0.002
Concatenation	0.732 ± 0.004	0.741 ± 0.004
Highway network fusion	$\mathbf{0.735 \pm 0.003}$	$\mathbf{0.747 \pm 0.002}$
w/o Deep Fusion module	0.727 ± 0.010	0.738 ± 0.007

As illustrated in Table 2, we conduct an in-depth sensitivity analysis on the number of Transformer layers across different modules. Our findings indicate that a configuration with 1-layer text encoder, 1-layer cross-modal attention, and 2-layer deep fusion achieves optimal results, balancing model capacity and overfitting risk. These observations align with the intuition that deeper fusion layers help the model capture complex cross-modal dependencies effectively.

Table 2. Impact of transformer depth on MUSE multilevel transformer model.

Text Encoder	**Cross-Mod**	**Deep Fusion**	**WA**	**UA**
3	3	1	0.723	0.734
2	2	1	0.730	0.739
1	1	1	0.731	0.743
1	1	2	0.735	**0.747**
1	1	3	0.726	0.734
2	2	2	**0.736**	0.744
2	2	3	0.725	0.733

4.4. Assessment of MUSE's Multigranular Fusion Framework

We further evaluate the full MUSE framework with multigranular fusion. As depicted in Table 3, integrating global utterance-level representations from BERT with fine-grained multimodal features yields a notable performance uplift, achieving a WA of 0.752 and UA of 0.756. This result demonstrates the clear advantage of our multigranular fusion strategy, which synergistically combines complementary information from different granularity levels. The empirical findings confirm that such unified fusion can effectively bridge the gap between local acoustic-textual patterns and global semantic cues, enhancing the robustness and discriminative power of emotion recognition systems.

Table 3. Performance comparison of MUSE multigranular model and its core components.

Methods	WA	UA
BERT (utterance-only)	0.693 ± 0.004	0.695 ± 0.001
MUSE Multilevel Transformer	0.735 ± 0.003	0.747 ± 0.002
MUSE Multigranular Fusion Model	$\mathbf{0.752 \pm 0.003}$	$\mathbf{0.756 \pm 0.006}$

5. Conclusions and Future Directions

In this work, we introduced **MUSE**, a novel Multigranular Unified Synthesis Encoder framework designed to advance the state-of-the-art in speech emotion recognition (SER). At the core of MUSE lies an innovative multilevel Transformer-based module, meticulously architected to enable fine-grained and hierarchical interactions across diverse modalities, specifically voice fragments, words, and phonemes. To the best of our knowledge, this constitutes the first endeavor to repurpose and adapt the structural principles of Transformer TTS [15] within the domain of SER. Our approach bridges the gap between coarse-grained sentence-level representations and fine-grained token-level features, facilitating more nuanced emotional understanding. Furthermore, we proposed a carefully designed multi-granularity fusion strategy within MUSE, which effectively amalgamates the fine-grained representations generated by our multilevel Transformer with global utterance-level embeddings derived from powerful pre-trained language models, such as BERT [4]. Unlike existing methods that either focus solely on pre-trained models or require costly alignment annotations, our framework offers a simple yet highly effective fusion mechanism, which significantly enhances the capability of multimodal emotion recognition systems.

Extensive and rigorous experimental evaluations conducted on the benchmark IEMOCAP dataset [17] demonstrate that our proposed MUSE framework consistently outperforms previous state-of-the-art methods across all evaluated metrics, including Weighted Accuracy (WA) and Unweighted Accuracy (UA). In particular, the multilevel Transformer component alone surpasses prior multimodal fusion approaches, while the complete multigranularity fusion framework further elevates performance, indicating the complementary nature of fine-grained and pre-trained representations. Our ablation studies also verify the robustness and critical contributions of each module within MUSE.

We believe that our proposed MUSE framework not only presents an effective solution for SER but also offers broader implications for other multimodal understanding tasks. By demonstrating the feasibility and superiority of integrating fine-grained cross-modal interactions with global semantic embeddings, our approach can serve as a reference design for future works involving other modalities or tasks, such as sentiment analysis, empathy modeling, or multimodal dialogue systems. To promote further research and facilitate reproducibility, we will release our code, pretrained models, and detailed training configurations to the public community. Looking forward, we envisage several promising avenues for extending this work. One natural direction is to incorporate recent advances in self-supervised acoustic models, such as Wav2Vec 2.0 [8], into the MUSE framework. By replacing or complementing the current mel spectrogram representation with more expressive representations from such pre-trained acoustic models, we expect to further enhance the emotion recognition capability, particularly in low-resource and noisy scenarios. Additionally, exploring adaptive granularity fusion strategies, where the fusion weightings are dynamically adjusted based on context or emotional intensity, could further refine our framework's effectiveness.

Another research direction involves extending MUSE into a broader multi-task learning paradigm, jointly optimizing emotion recognition along with auxiliary tasks such as speaker identification, sentiment detection, or affective reasoning. We hypothesize that such auxiliary tasks could provide beneficial supervision signals to enrich the emotion understanding capability of the model. Finally, applying and validating MUSE on more diverse datasets, including in-the-wild conversational datasets and multi-language emotional corpora, will be critical to assess the generalization potential of our proposed methods. In conclusion, we hope our work serves as a stepping stone toward more general,

 doi:10.20944/preprints202505.1230.v1

robust, and explainable multimodal emotion understanding systems, and we encourage the community to further explore and extend the directions outlined in this paper.

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.
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. 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).
33. 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.
34. 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.
35. 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)*.
36. 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.
37. Ilya Loshchilov and Frank Hutter. 2019. Decoupled Weight Decay Regularization. In *International Conference on Learning Representations*.
38. 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).
39. Pradeep K. Atrey, M. Anwar Hossain, Abdulmotaleb El Saddik, and Mohan S. Kankanhalli. Multimodal fusion for multimedia analysis: a survey. *Multimedia Systems*, 16(6):345–379, April 2010. ISSN 0942-4962. 10.1007/s00530-010-0182-0.
40. Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, may 2015. URL http://dx.doi.org/10.1038/nature14539.
41. 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/.
42. Eric Makita and Artem Lenskiy. A movie genre prediction based on Multivariate Bernoulli model and genre correlations. (May), mar 2016. URL http://arxiv.org/abs/1604.08608.
43. Junhua Mao, Wei Xu, Yi Yang, Jiang Wang, and Alan L Yuille. Explain images with multimodal recurrent neural networks. *arXiv preprint arXiv:1410.1090*, 2014.
44. Deli Pei, Huaping Liu, Yulong Liu, and Fuchun Sun. Unsupervised multimodal feature learning for semantic image segmentation. In *The 2013 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–6. IEEE, aug 2013. ISBN 978-1-4673-6129-3. 10.1109/IJCNN.2013.6706748. URL http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6706748.
45. Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.
46. Richard Socher, Milind Ganjoo, Christopher D Manning, and Andrew Ng. Zero-Shot Learning Through Cross-Modal Transfer. In C J C Burges, L Bottou, M Welling, Z Ghahramani, and K Q Weinberger (eds.),

 doi:10.20944/preprints202505.1230.v1

Advances in Neural Information Processing Systems 26, pp. 935–943. Curran Associates, Inc., 2013. URL http://papers.nips.cc/paper/5027-zero-shot-learning-through-cross-modal-transfer.pdf.

47. 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.
48. A. Karpathy and L. Fei-Fei, "Deep visual-semantic alignments for generating image descriptions," *TPAMI*, vol. 39, no. 4, pp. 664–676, 2017.
49. 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.
50. 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.
51. 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.
52. 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.
53. 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.
54. Ling Zhuang, Hao Fei, and Po Hu. Knowledge-enhanced event relation extraction via event ontology prompt. *Inf. Fusion*, 100:101919, 2023.
55. 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.
56. 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.
57. 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.
58. 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.
59. 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.
60. 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.
61. 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.
62. 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.
63. Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*, 2016.
64. 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.
65. D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *ICLR*, 2015.
66. 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.
67. K. Papineni, S. Roukos, T. Ward, and W. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *ACL*, 2002, pp. 311–318.
68. 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.

 doi:10.20944/preprints202505.1230.v1

69. 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.
70. Shengqiong Wu, Hao Fei, Leigang Qu, Wei Ji, and Tat-Seng Chua. Next-gpt: Any-to-any multimodal llm. *CoRR*, abs/2309.05519, 2023.
71. 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.
72. 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.
73. Naman Jain, Pranjali Jain, Pratik Kayal, Jayakrishna Sahit, Soham Pachpande, Jayesh Choudhari, et al. Agribot: agriculture-specific question answer system. *IndiaRxiv*, 2019.
74. 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.
75. 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.
76. 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.
77. 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.
78. 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.
79. 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.
80. S. Banerjee and A. Lavie, "METEOR: an automatic metric for MT evaluation with improved correlation with human judgments," in *IEEMMT*, 2005, pp. 65–72.
81. 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.
82. 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.
83. 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.
84. 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.
85. 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.
86. 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.
87. 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.
88. 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.

doi:10.20944/preprints202505.1230.v1

89. 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.
90. 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.
91. 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.
92. P. Anderson, B. Fernando, M. Johnson, and S. Gould, "SPICE: semantic propositional image caption evaluation," in *ECCV*, 2016, pp. 382–398.
93. 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.
94. 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.
95. 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.
96. 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.
97. Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhutdinov, Richard S Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. *arXiv preprint arXiv:1502.03044*, 2(3):5, 2015.
98. Seniha Esen Yuksel, Joseph N Wilson, and Paul D Gader. Twenty years of mixture of experts. *IEEE transactions on neural networks and learning systems*, 23(8):1177–1193, 2012.
99. Sanjeev Arora, Yingyu Liang, and Tengyu Ma. A simple but tough-to-beat baseline for sentence embeddings. In *ICLR*, 2017.