Article

Transformer-based Visual Expression Identification with Recurrent Neural Network

Noah Brown , Emily Marwood , Liam Smith , Olivia Williams *

Posted Date: 18 March 2025

doi: 10.20944/preprints202503.1265.v1

Keywords: expression recognition; multimodal transformers; human-computer interaction

Article

Transformer-Based Visual Expression Identification with Recurrent Neural Network

Noah Brown, Emily Marwood, Liam Smith and Olivia Williams *

Bond University
* Correspondence: owilliams@bond.edu.au

Abstract: Creating sophisticated machine learning models to comprehend interactions between individuals can lead to more intuitive user experiences for interactive systems like Amazon Alexa. Beyond basic indicators such as voice modulation and eye movement, a person's combined audio-visual expressions—including vocal intonation and facial gestures—act as subtle cues reflecting the level of engagement in a conversation. This research explores advanced deep learning techniques for the detection of user expressions through audio-visual data. Initially, we develop a foundational audio-visual model incorporating recurrent neural network layers, which demonstrates performance on par with existing leading methods. Subsequently, we introduce a novel transformer-based framework equipped with encoder layers that more effectively fuse audio and visual features for tracking expressions. Evaluation using the Aff-Wild2 dataset reveals that our proposed transformer models outperform the recurrent-based baseline by approximately 2% in accurately identifying arousal and valence metrics. Additionally, our multimodal transformer approaches exhibit notable enhancements compared to unimodal models, achieving performance improvements of up to 3.6%. Comprehensive ablation analyses confirm the crucial role of visual information in the accurate detection of expressions within the Aff-Wild2 dataset. These findings underscore the potential of transformer architectures in advancing the field of expression recognition and enhancing human-computer interaction systems.

Keywords: expression recognition; multimodal transformers; human-computer interaction

1. Introduction

Audio-visual communication stands as a cornerstone among the various modes of human interaction. Humans possess an innate ability to perceive and respond to their social environments, which is often expressed through external displays of emotion and expression. These expressions are integral to effective communication, facilitating a deeper understanding between individuals.

With the rapid advancement and widespread adoption of interactive devices such as Amazon Alexa, Google Home, and other smart home technologies, there is an increasing need for these systems to interpret and respond to user emotions accurately. The ability to detect and monitor user expressions, encompassing both affective states and spoken language, is pivotal in creating more intuitive and natural user interfaces. These smart devices are increasingly embedded in diverse sectors, including communication platforms, healthcare monitoring systems [1], security infrastructures [2], and educational tools [3], underscoring the multifaceted applications of emotion detection technologies.

Human communication extends beyond verbal interactions. In addition to speech [5,69], individuals utilize a plethora of non-verbal cues such as body posture, gestures, facial expressions, and eye movements to convey emotions and intentions. Consequently, recognition algorithms have evolved to prioritize these explicit forms of expression. Research has extensively explored variations in vocal tone and speech patterns [6], detailed facial expressions [7,8,70,77], as well as body language and posture [9,10]. Furthermore, physiological indicators like heart rate variability [11,12] have been investigated to provide a comprehensive understanding of an individual's emotional state. Among these modalities, vocal tone (audio) and facial expressions (video) remain the most prominent channels for communicating affect [13]. The ubiquity of audio and video sensors in modern devices enhances their appeal for developing robust expression detection systems.

The task of expression detection is primarily categorized based on the nature of labeling techniques employed. One prevalent approach is discrete classification, which assigns expressions to a limited set of predefined categories such as happiness, sadness, anger, and neutrality [6]. This method simplifies the continuum of human emotions into distinct classes, facilitating straightforward classification tasks. Alternatively, descriptor-based methods quantify expressions along continuous dimensions, most commonly arousal (intensity of emotion) and valence (positivity or negativity) [4]. This dimensional approach allows for a more nuanced representation of emotions, capturing the subtleties that discrete categories may overlook. Recently, there has been a growing interest in predicting specific attribute values to describe expressions with greater granularity [14], enabling the detection of more nuanced emotional states.

In this study, we focus on the multimodal detection of arousal and valence attributes, emphasizing the integration of visual and auditory modalities. We introduce a novel transformer-based architecture, named **ExpTrm**, which incorporates cross-modal attention mechanisms to effectively combine audio and visual features [7,87]. The **ExpTrm** model leverages self-attention encoder layers to identify and prioritize significant cues within each modality while simultaneously utilizing cross-modal attention layers to integrate information across modalities. This dual attention mechanism facilitates the capture of temporal and contextual dependencies within audio-visual clips, enhancing the model's ability to accurately detect and interpret expressions.

To evaluate the efficacy of the **ExpTrm** architecture, we conduct experiments using the Aff-Wild2 database [16,89], a comprehensive dataset for affective computing. Our evaluation focuses on predicting arousal and valence scores, which are critical for understanding the intensity and nature of user emotions. Initially, we establish a competitive baseline model employing recurrent neural network (RNN) layers to capture temporal dynamics in audio and video signals. This baseline demonstrates superior performance compared to existing models that utilize single modalities, highlighting the importance of temporal context in expression detection.

Subsequently, we demonstrate that the **ExpTrm** model significantly outperforms the RNN-based baseline, achieving absolute improvements of 1.7% in valence and 1.9% in arousal metrics. These enhancements indicate the superiority of the transformer-based architecture in handling the complexities of multimodal expression detection tasks, establishing new benchmarks on the Aff-Wild2 dataset. Furthermore, our multimodal **ExpTrm** models exhibit substantial performance gains over unimodal approaches, with improvements reaching up to 3.6%, underscoring the benefits of integrating multiple sensory inputs for more accurate emotion recognition.

Additionally, we perform comprehensive ablation studies to dissect the contributions of each modality within the **ExpTrm** framework. These studies reveal that the visual modality plays a more dominant role in recognizing expressions, although the auditory modality also contributes significantly to the overall performance. This insight highlights the potential for further optimizing multimodal fusion strategies to enhance emotion detection accuracy.

In summary, this paper makes the following key contributions: 1) Development of a robust recurrent neural network (RNN) baseline model for single and multimodal expression detection tasks, demonstrating competitive performance in capturing temporal dynamics. 2) Introduction of the **ExpTrm**, a multimodal transformer architecture featuring cross-modal attention layers for effective fusion of audio and visual modalities, achieving state-of-the-art results in expression recognition. These contributions advance the field of affective computing by providing more accurate and reliable methods for emotion detection, thereby facilitating the creation of more responsive and empathetic human-computer interaction systems.

2. Related Work

Audio Cues

Extensive research has been conducted on emotion recognition utilizing audio cues. Traditional approaches have predominantly focused on two main paradigms: categorical emotion recognition, which classifies discrete utterances into predefined emotional states [17,91], and continuous emotion prediction, which estimates emotional states on a continuous scale [18,102]. A diverse array of acoustic features is employed to facilitate accurate emotion detection. Commonly used features include log-spectrograms and mel-frequency filterbanks, which capture the spectral properties of audio signals [19,93]. Additionally, para-linguistic features, which encompass aspects such as speech rate, pitch variation, and intensity, have been extensively explored for their efficacy in emotion recognition tasks [20,21]. These features are typically extracted from short, overlapping frames to effectively describe the dynamic changes in vocal effort associated with varying emotional expressions. For a comprehensive overview of architectural frameworks tailored for audio-based expression detection, Zeng *et al.* provide an in-depth analysis [22]. Recent advancements have also seen the integration of deep learning techniques, which leverage convolutional and recurrent neural networks to automatically learn and extract high-level audio features, thereby enhancing the robustness and accuracy of emotion recognition systems.

Facial Expressions

Facial expression analysis has been a focal point in the domain of affective computing, with numerous studies dedicated to decoding emotions from visual cues. Facial expressions are typically categorized into discrete emotional states or described in terms of facial action units, which correspond to specific muscle movements [22]. Detection methodologies vary based on the task at hand, encompassing both static image recognition [23,24,99] and continuous prediction across image sequences [25]. Traditional approaches often begin with face alignment techniques to isolate the facial region from the surrounding background, ensuring that subsequent analysis focuses solely on relevant facial features. Geometric visual cues, such as the shape and configuration of facial landmarks (e.g., eyes, mouth, eyebrows) and their spatial relationships (e.g., distances between the corners of the eyes and mouth), are extensively utilized for recognizing facial expressions [26,27,78]. Landmark detection algorithms are typically employed to accurately extract these visual features [28]. With the advent of deep learning, there has been a significant shift towards data-driven models that operate on large-scale, in-the-wild datasets, thereby improving the generalizability and robustness of facial expression recognition systems [28,29]. Similar to these prior studies, our research employs a sophisticated face detection mechanism to extract facial regions, which are then processed using a pre-trained deep learning model to capture intricate facial features.

Audio-Visual Expression Detection

The integration of audio and visual modalities for expression detection has garnered substantial attention in recent studies. Researchers have developed various machine learning models aimed at predicting audiovisual expressions by leveraging the complementary nature of audio and visual information [30–32,107,108]. It has been demonstrated that while audio cues are particularly effective in predicting arousal levels—where changes in vocal tone correlate with fluctuations in emotional intensity—visual cues are more adept at discerning valence, indicating the positivity or negativity of the emotion. The fusion of these modalities can be achieved through different strategies, including feature-level (early) fusion, where audio and visual features are combined before model processing; decision-level (late) fusion, where predictions from separate audio and visual models are aggregated; or hybrid fusion approaches that incorporate elements of both [22]. The rise of deep learning has facilitated more sophisticated early fusion techniques, enabling data-driven models to jointly learn and integrate multimodal features [33]. Building upon these advancements, our work introduces an innovative attention-based mechanism to effectively merge latent audio-visual features, thereby enhancing the accuracy of expression detection.

Attention Mechanisms for Expression Recognition

Attention mechanisms have revolutionized the way models process and prioritize information within signals, playing a pivotal role in enhancing the performance of emotion recognition systems. Fundamentally, attention mechanisms allocate computational resources to the most relevant parts of the input data, thereby improving the model's focus and interpretability. In the context of audio cues, Chen *et al.* [34,110] proposed a three-dimensional convolutional recurrent neural network augmented with an attention mechanism that selectively emphasizes significant speech frames, resulting in improved unweighted recall rates for emotion classification tasks. For visual cues, Xiaohua *et al.* [35,86] introduced a two-stage attention network designed to model the interrelationships between various positional features on the face, thereby capturing subtle emotional nuances. Hazarika *et al.* [36,85] employed a soft attention mechanism to effectively capture and retain relevant information from audio, visual, and textual modalities during dyadic interactions, enhancing the model's ability to understand conversational dynamics.

Traditionally, recurrent layers such as LSTMs and GRUs have been the backbone for capturing temporal dependencies in sequential data for emotion prediction. Mirsamadi *et al.* [37] incorporated a local attention mechanism atop recurrent layers to selectively pool emotion-specific audio features, thereby refining the emotion recognition process. More recently, there has been a shift towards utilizing self-attention mechanisms as an alternative to recurrent architectures, aiming to capture long-range dependencies more efficiently. Li *et al.* [38] introduced a self-attention module integrated with convolutional and recurrent layers, which significantly boosted emotion recognition performance from speech data. Rahman *et al.* [39] explored methods to integrate audio and visual cues into transformers that were initially pretrained on textual data, using audio-visual information to modulate and gate textual features for enhanced emotion recognition.

In contrast to the aforementioned studies, our work presents a novel cross-modal attention mechanism within the **ExpTrm** framework, specifically designed to fuse audio and visual cues for more accurate expression detection. This cross-modal attention facilitates the dynamic interaction and integration of information across different modalities, enabling the model to prioritize and leverage the most salient features from both audio and visual inputs. By doing so, **ExpTrm** not only enhances the model's ability to capture complex emotional states but also sets a new benchmark in the field of multimodal emotion recognition.

3. Methodology

This study focuses on the continuous recognition of expressions using a novel multimodal transformer-based architecture, termed **ExpTrm**. The primary objective is to accurately predict the arousal and valence values for each frame within a sequence of audio-visual data. Given the complexity and richness of human expressions, leveraging both audio and visual modalities provides a comprehensive understanding of the underlying emotional states.

3.1. Overall Framework

The fusion of audio and video streams can be approached through various methodologies. An intuitive and data-driven strategy is to enable the model to autonomously learn the integration of these modalities. The attention mechanism emerges as a particularly promising framework for this purpose, facilitating the dynamic weighting of relevant features from each modality. Building upon the foundational work of Vaswani *et al.* [15], which demonstrated the efficacy of dot-product attention in achieving state-of-the-art results across diverse tasks [40], we propose the **ExpTrm** model. This model incorporates cross-modal attention layers that utilize dot-product attention mechanisms to learn robust feature representations conducive to accurate expression prediction.

Cross-modal attention has proven effective in various applications, such as audio-visual automatic speech recognition [41]. The rationale behind employing cross-modal attention in **ExpTrm** is to ensure that the learned features are resilient to potential occlusions or missing data in either modality. For instance, if a user's face is not within the camera's field of view (FoV), the model can still infer the user's expression based on audio cues. Conversely, when audio information is scarce or absent, visual cues can compensate to maintain prediction accuracy [42].

The architecture of **ExpTrm** comprises two primary encoders: one dedicated to processing audio data and the other to visual data. Each encoder is built upon multiple self-attention layers, which individually attend to significant features within their respective modalities. The embeddings generated by these encoders are subsequently fused through two cross-modal attention layers, facilitating the integration of information across modalities. The final fused features are then utilized to predict the arousal and valence values for each frame.

3.2. Transformer Architecture

The transformer architecture, introduced by Vaswani *et al.* [15], has revolutionized the field of deep learning by enabling efficient and scalable modeling of sequential data. Unlike traditional recurrent neural networks (RNNs), transformers leverage self-attention mechanisms to capture dependencies between input elements, regardless of their positional distance in the sequence. This section provides a detailed overview of the transformer architecture, incorporating essential formulas to elucidate its operational principles.

3.2.1. Self-Attention Mechanism

At the core of the transformer architecture lies the self-attention mechanism, which allows the model to weigh the importance of different parts of the input sequence when encoding a particular element. Given an input sequence of vectors $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n]$, the self-attention mechanism computes three distinct matrices: queries ($\mathbf{Q}$), keys ($\mathbf{K}$), and values ($\mathbf{V}$).

$$\mathbf{Q} = \mathbf{X}\mathbf{W}^Q, \quad \mathbf{K} = \mathbf{X}\mathbf{W}^K, \quad \mathbf{V} = \mathbf{X}\mathbf{W}^V \tag{1}$$

where $\mathbf{W}^Q, \mathbf{W}^K, \mathbf{W}^V$ are learned weight matrices for queries, keys, and values, respectively.

The attention scores are computed using the scaled dot-product attention formula:

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d_k}}\right)\mathbf{V} \tag{2}$$

where d_k is the dimensionality of the keys. The softmax function ensures that the attention weights sum to one, effectively highlighting the most relevant parts of the input sequence.

3.2.2. Multi-Head Attention

To capture diverse aspects of the input data, transformers employ multi-head attention, which involves multiple parallel attention layers, or "heads." Each head operates independently, allowing the model to attend to different representation subspaces at different positions.

$$\text{MultiHead}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{Concat}(\text{head}_1, \ldots, \text{head}_h)\mathbf{W}^O \tag{3}$$

where each head head_i is computed as:

$$\text{head}_i = \text{Attention}(\mathbf{Q}\mathbf{W}_i^Q, \mathbf{K}\mathbf{W}_i^K, \mathbf{V}\mathbf{W}_i^V) \tag{4}$$

and $\mathbf{W}^O$ is the output weight matrix. This mechanism enhances the model's ability to focus on different parts of the input simultaneously, thereby enriching the learned representations.

3.2.3. Positional Encoding

Since transformers do not inherently capture the sequential order of input data, positional encodings are added to the input embeddings to inject information about the position of each element in the sequence. These encodings are typically sinusoidal functions of different frequencies, defined as:

$$\mathrm{PE}(pos, 2i) = \sin\left(\frac{pos}{10000^{\frac{2i}{d_{model}}}}\right) \tag{5}$$

$$\mathrm{PE}(pos, 2i+1) = \cos\left(\frac{pos}{10000^{\frac{2i}{d_{model}}}}\right) \tag{6}$$

where pos is the position and i is the dimension. These encodings enable the model to discern the order of the sequence, which is crucial for tasks involving sequential data.

3.3. ExpTrm Architecture

The **ExpTrm** model extends the traditional transformer architecture by incorporating cross-modal attention mechanisms to effectively fuse audio and visual features. The architecture is meticulously designed to handle the distinct characteristics of each modality while ensuring seamless integration for robust expression detection.

3.3.1. Modal Encoders

ExpTrm employs two separate encoders: one for audio and one for visual data. Each encoder consists of multiple self-attention layers, enabling the model to capture intricate patterns and dependencies within each modality independently.

Audio Encoder

The audio encoder processes the audio stream, extracting relevant temporal and spectral features. The input audio frames are first transformed into log-spectrograms or mel-frequency filterbanks, which serve as the input embeddings. These embeddings are then passed through a series of self-attention layers, allowing the model to focus on significant audio cues indicative of arousal and valence.

Visual Encoder

The visual encoder handles the video stream, focusing on facial expressions and other visual cues. Facial regions are extracted using a pre-trained face detector, and the cropped face images are processed through a convolutional neural network (CNN) to obtain high-level visual features. These features are subsequently fed into the self-attention layers of the visual encoder, enabling the model to capture spatial and temporal dynamics of facial expressions.

3.3.2. Cross-Modal Attention Layers

After processing the audio and visual streams independently, the resulting embeddings are fused through two cross-modal attention layers. These layers employ the dot-product attention mechanism, similar to the self-attention layers, but with a crucial modification: the key ($\mathbf{K}$) and value ($\mathbf{V}$) matrices are derived from one modality, while the query ($\mathbf{Q}$) matrix is derived from the opposite modality.

Formally, for the audio-to-visual cross-modal attention:

$$\mathbf{Q}_{\text{audio}} = \mathbf{H}_{\text{audio}}\mathbf{W}^{Q}, \quad \mathbf{K}_{\text{visual}} = \mathbf{H}_{\text{visual}}\mathbf{W}^{K}, \quad \mathbf{V}_{\text{visual}} = \mathbf{H}_{\text{visual}}\mathbf{W}^{V} \tag{7}$$

$$\text{Attention}_{\text{audio-visual}} = \text{softmax}\left(\frac{\mathbf{Q}_{\text{audio}}\mathbf{K}_{\text{visual}}^{\top}}{\sqrt{d_k}}\right)\mathbf{V}_{\text{visual}} \tag{8}$$

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 March 2025 doi:10.20944/preprints202503.1265.v1

Similarly, for the visual-to-audio cross-modal attention:

$$\mathbf{Q}_{\text{visual}} = \mathbf{H}_{\text{visual}}\mathbf{W}^{Q}, \quad \mathbf{K}_{\text{audio}} = \mathbf{H}_{\text{audio}}\mathbf{W}^{K}, \quad \mathbf{V}_{\text{audio}} = \mathbf{H}_{\text{audio}}\mathbf{W}^{V} \tag{9}$$

$$\text{Attention}_{\text{visual-audio}} = \text{softmax}\left(\frac{\mathbf{Q}_{\text{visual}}\mathbf{K}_{\text{audio}}^{\top}}{\sqrt{d_k}}\right)\mathbf{V}_{\text{audio}} \tag{10}$$

These cross-modal attention layers enable the model to dynamically integrate information from both audio and visual modalities, allowing **ExpTrm** to prioritize relevant features from each modality based on the contextual interplay between them.

3.3.3. Feature Fusion and Prediction

The outputs from the cross-modal attention layers are combined using a weighted sum approach, where scalar weights α and β are learned parameters that balance the contributions from each modality:

$$\mathbf{F}_{\text{fused}} = \alpha \cdot \text{Attention}_{\text{audio-visual}} + \beta \cdot \text{Attention}_{\text{visual-audio}} \tag{11}$$

This fused feature representation encapsulates the integrated information from both audio and visual streams, enriched by the cross-modal interactions facilitated by the attention mechanisms.

Subsequently, the fused features are passed through a fully connected layer, followed by non-linear activation functions, to regress the arousal and valence values for each frame. The prediction layer is defined as:

$$\mathbf{y} = \text{ReLU}(\mathbf{F}_{\text{fused}}\mathbf{W} + \mathbf{b}) \tag{12}$$

where **W** and **b** are learnable parameters, and **y** represents the predicted arousal and valence values.

3.4. Training Procedure

The **ExpTrm** model is trained end-to-end using the Aff-Wild2 database [16], which provides a comprehensive set of annotated audio-visual data for affective computing tasks. The training objective is to minimize the mean squared error (MSE) between the predicted and ground truth arousal and valence values:

$$\mathcal{L} = \frac{1}{N}\sum_{i=1}^{N}\left((\hat{a}_i - a_i)^2 + (\hat{v}_i - v_i)^2\right) \tag{13}$$

where N is the number of training samples, $\hat{a}_i$ and $\hat{v}_i$ are the predicted arousal and valence values, and a_i and v_i are the corresponding ground truth values.

To enhance the model's generalization capabilities, various regularization techniques such as dropout and weight decay are employed. Additionally, early stopping based on validation loss is implemented to prevent overfitting.

3.5. Handling Missing Modalities

One of the inherent challenges in multimodal expression detection is the potential absence or occlusion of one of the modalities. The **ExpTrm** architecture is designed to be resilient to such scenarios through its cross-modal attention mechanism. When one modality is missing or unreliable, the attention layers can dynamically adjust to rely more heavily on the available modality, ensuring that expression predictions remain robust and accurate.

For instance, if the visual input is occluded, the model can leverage the audio stream to infer the user's emotional state based on vocal cues. Conversely, in the absence of audio data, the model can rely on visual features to maintain prediction performance. This adaptability is crucial for real-world applications where perfect multimodal data is often unattainable.

3.6. Remarks on ExpTrm

The **ExpTrm** model offers several advantages over traditional multimodal fusion approaches:

- **Dynamic Feature Integration:** The cross-modal attention layers enable dynamic weighting and integration of audio and visual features, allowing the model to prioritize relevant information based on the contextual interplay between modalities.
- **Scalability:** Leveraging the transformer architecture facilitates scalability to longer sequences and larger datasets, enhancing the model's ability to capture complex temporal dependencies.
- **Robustness to Missing Data:** The inherent design of cross-modal attention provides resilience against missing or occluded modalities, ensuring consistent performance in diverse scenarios.
- **State-of-the-Art Performance:** Empirical evaluations demonstrate that **ExpTrm** achieves superior performance in predicting arousal and valence values, setting new benchmarks on the Aff-Wild2 dataset.

While **ExpTrm** demonstrates significant advancements in multimodal expression detection, several avenues for future research remain:

- **Incorporation of Additional Modalities:** Integrating other sensory inputs, such as physiological signals (e.g., heart rate, skin conductance), could further enhance emotion recognition accuracy.
- **Real-Time Processing:** Optimizing the model for real-time expression detection would expand its applicability in interactive systems and live monitoring scenarios.
- **Personalization:** Developing personalized models that adapt to individual differences in emotional expression could improve performance in diverse user populations.
- **Robustness to Adversarial Conditions:** Enhancing the model's resilience to noisy or adversarial inputs would ensure reliable performance in real-world environments.

These directions aim to build upon the foundations established by **ExpTrm**, pushing the boundaries of multimodal emotion recognition and its applications in human-computer interaction.

4. Experimental Setup

4.1. Dataset

In this study, we employ the Aff-Wild2 database [16], which stands out among existing datasets for its extensive collection of in-the-wild recordings. Unlike other datasets that are often limited to controlled environments, Aff-Wild2 offers a diverse range of real-world scenarios, capturing subjects under varying conditions such as different lighting, backgrounds, and levels of expressiveness. This diversity is crucial as it mirrors the complexities encountered in practical applications, making it an ideal benchmark for evaluating the fusion of audio and video cues—a central focus of this research. The dataset comprises 558 videos, predominantly featuring individual subjects engaged in self-expression. These videos are recorded at a consistent frame rate of 30 frames per second (fps), accumulating to approximately 2.8 million video frames in total.

Aff-Wild2 is meticulously annotated for three primary behavioral tasks: basic expression classification, valence-arousal estimation, and action unit detection. For the purpose of this study, we concentrate on frame-level valence-arousal estimation, leveraging the continuous annotations provided. Each frame in the dataset is annotated with valence and arousal values on a continuous scale ranging from -1 to 1, capturing the intensity and positivity/negativity of the expressed emotion. These annotations are sourced from the consensus of 4 to 8 annotators per frame, ensuring high reliability and consistency in the labeling process. To facilitate robust training and unbiased evaluation, the dataset is partitioned into speaker-independent subsets: 350 videos for training, 70 for validation, and 138 for testing. This split ensures that the model's performance is evaluated on unseen subjects, thereby assessing its generalizability across diverse individuals and expression styles.

4.2. Features and Preprocessing

Video

For the visual modality, our approach begins with the extraction and preprocessing of facial features from video frames. Following established methodologies [43], we utilize a Single Shot MultiBox Detector (SSD) with a ResNet backbone [44] to accurately detect and localize frontal faces within each frame. This detector is adept at handling variations in face orientation, scale, and lighting conditions, ensuring robust face detection across the diverse conditions present in the Aff-Wild2 dataset. Once a face is detected, the corresponding region is cropped from the frame and resized to a standardized dimension of 96x96 pixels with 3 color channels (96x96x3). This resizing ensures uniformity across all inputs, facilitating efficient processing by subsequent neural network layers. Additionally, all pixel intensities are normalized to fall within the range of [-1, 1], a common preprocessing step that aids in stabilizing the training process and accelerating convergence by maintaining consistent input distributions.

To further enhance the quality of the extracted facial features, we employ a pre-trained VGGFace network [46]. Specifically, we extract descriptors from the first fully connected (FC) layer of the network, which contains 4096 nodes. These descriptors encapsulate high-level facial features, capturing intricate details that are critical for accurate emotion recognition. By leveraging a pre-trained network, we benefit from transfer learning, where the model's learned representations from large-scale face datasets enhance our ability to extract meaningful features from the Aff-Wild2 data. These visual features serve as rich representations of the facial expressions, providing a robust foundation for the transformer-based architecture to analyze and predict emotional states.

Audio

For the auditory modality, we extract a comprehensive set of features tailored for emotion recognition. Specifically, we utilize the feature set introduced for the para-linguistic challenge at Interspeech 2013 [45], which has demonstrated state-of-the-art performance across various emotion recognition tasks. This feature set comprises *Low-Level Descriptors* (LLDs) and *High-Level Descriptors* (HLDs). LLDs are extracted over short temporal windows of 25 milliseconds with a 10-millisecond step size, capturing fine-grained variations in the audio signal. These descriptors include spectral features such as Mel Frequency Cepstral Coefficients (MFCCs), which represent the short-term power spectrum of sound; energy-related features like loudness, which indicate the intensity of the audio signal; and voice-based features such as the fundamental frequency (F0), which correlates with the perceived pitch of the voice.

While HLDs, which are statistical aggregations of LLDs over larger segments, offer more comprehensive representations, they result in high-dimensional feature vectors (6373 dimensions), posing computational challenges and the risk of overfitting. To address this, we opt to utilize only the 65-dimensional LLDs, which strike a balance between feature richness and computational efficiency. These LLDs encompass a broad spectrum of auditory characteristics essential for emotion detection. Following extraction, the audio features undergo Z-normalization based on the statistics derived from the training set. This normalization process ensures that each feature has zero mean and unit variance, promoting numerical stability and enhancing the model's ability to learn effectively from the data.

4.3. Transformer Model Architecture

At the heart of our proposed methodology lies the transformer-based architecture, designated as **ExpTrm**. Building upon the foundational work of Vaswani *et al.* [15], transformers leverage self-attention mechanisms to capture complex dependencies within sequential data, eschewing the need for recurrent structures inherent in traditional models. This attribute makes transformers particularly well-suited for tasks involving temporal sequences, such as those encountered in audio and video data streams.

The **ExpTrm** architecture comprises two primary encoders: one dedicated to processing audio inputs and the other to visual inputs. Each encoder is constructed from multiple self-attention layers, enabling the model to focus on salient features within each modality independently. The self-attention mechanism operates by generating three distinct matrices—queries (**Q**), keys (**K**), and values (**V**)—from the input embeddings through learned linear transformations:

$$\mathbf{Q} = \mathbf{X}\mathbf{W}^Q, \quad \mathbf{K} = \mathbf{X}\mathbf{W}^K, \quad \mathbf{V} = \mathbf{X}\mathbf{W}^V \tag{14}$$

doi:10.20944/preprints202503.1265.v1

where $\mathbf{W}^Q, \mathbf{W}^K, \mathbf{W}^V$ are trainable weight matrices specific to queries, keys, and values, respectively. The attention scores are computed using the scaled dot-product attention formula:

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{QK}^\top}{\sqrt{d_k}}\right)\mathbf{V} \tag{15}$$

Here, d_k represents the dimensionality of the key vectors, and the softmax function ensures that the attention weights sum to one, effectively highlighting the most relevant parts of the input sequence.

To capture diverse aspects of the input data, **ExpTrm** employs multi-head attention, which involves multiple parallel attention layers or "heads." Each head operates independently, allowing the model to attend to different representation subspaces at various positions within the input:

$$\text{MultiHead}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{Concat}(\text{head}_1, \ldots, \text{head}_h)\mathbf{W}^O \tag{16}$$

where each head head_i is computed as:

$$\text{head}_i = \text{Attention}(\mathbf{QW}_i^Q, \mathbf{KW}_i^K, \mathbf{VW}_i^V) \tag{17}$$

and $\mathbf{W}^O$ is the output weight matrix. This multi-head configuration allows **ExpTrm** to capture a wide array of feature interactions and dependencies, enriching the learned representations and enhancing the model's capacity to discern subtle emotional cues.

4.4. Architecture and Implementation

The overall architecture of **ExpTrm** is meticulously designed to facilitate the seamless integration of audio and visual modalities through advanced attention mechanisms. The architecture is structured as follows:

- **Audio and Video Encoders**: The model comprises two separate encoders—one for audio and one for video. Each encoder consists of multiple self-attention layers, enabling the extraction of relevant features within each modality independently.
- **Cross-Modal Attention Layers**: Following the individual encoders, the architecture incorporates two cross-modal attention layers. These layers utilize dot-product attention mechanisms to facilitate the interaction between audio and visual features, allowing the model to dynamically prioritize and integrate information from both modalities.
- **Feature Fusion**: The outputs from the cross-modal attention layers are combined using a weighted sum approach, where scalar trainable parameters α and β determine the contribution of each modality to the final fused feature representation.
- **Prediction Layer**: The fused features are fed into a dense output layer, which regresses the continuous valence and arousal values for each frame, enabling precise emotion detection.

Predictions are generated on sequences of length 100 frames, corresponding to approximately 3 seconds of video footage. To ensure synchronization between audio and video inputs, audio frames are downsampled by a factor of 3.3, aligning them temporally with the video frames. Additionally, the audio LLDs are concatenated over a two-second window to match the 4096-dimensional feature space of the video descriptors extracted from the VGGFace network. This synchronization ensures that audio and visual features are temporally aligned, facilitating effective multimodal fusion.

The concordance correlation coefficient (CCC) is employed as both the loss function and evaluation metric. CCC measures the agreement between predicted and ground truth values, accounting for both precision and accuracy. Mathematically, CCC is defined as:

$$\rho_c = \frac{2\rho\sigma_x\sigma_y}{\sigma_x^2 + \sigma_y^2 + (\mu_x - \mu_y)^2} \tag{18}$$

where ρ denotes Pearson's correlation coefficient, σ_x and σ_y represent the standard deviations of the predicted and ground truth values, respectively, and μ_x and μ_y are their respective means. By minimizing $1 - \rho_c$, the model is trained to maximize the concordance between predictions and true values.

The **ExpTrm** model is optimized using the Adam optimizer with an initial learning rate of 1×10^{-5}. To prevent overfitting and enhance generalization, techniques such as dropout and weight decay are employed. The training process is monitored using the validation set, and the best-performing model is selected based on the lowest validation loss, ensuring optimal performance on unseen data.

4.5. Baseline Models

To comprehensively evaluate the effectiveness of the proposed **ExpTrm** model, we establish a suite of baseline models derived from existing research and standard architectures. These baselines encompass both unimodal and multimodal approaches, providing a robust framework for comparative analysis. The baseline models are detailed as follows:

1 **AffWildNet + Static (V)**: As introduced by Kollias *et al.* [47], this baseline model is pretrained on the Aff-Wild database [47], the predecessor to Aff-Wild2, focusing exclusively on the video modality. The static variant processes each image independently using the VGGFace architecture, which comprises multiple convolutional layers followed by a fully connected (FC) layer and an output layer for predictions. We experiment with two configurations for the FC layer: one with 2000 dimensions and another with 4096 dimensions. It is important to note that our proposed **ExpTrm** model builds upon the VGGFace architecture used in this baseline, thereby sharing identical preprocessing steps and leveraging similar feature extraction mechanisms.

2 **AffWildNet + Dynamic (V)**: Extending the static approach, Kollias *et al.* [47] propose a dynamic model that processes sequences of images to capture temporal dynamics inherent in facial expressions. This dynamic variant maintains the same FC layer dimension of 4096 and incorporates two Gated Recurrent Unit (GRU) layers, each consisting of 128 nodes, to model temporal dependencies across the sequence. We adopt the same sequence length of 100 frames used in our transformer-based approach, ensuring consistency in temporal resolution and enabling a fair comparison between the models.

3 **RNN (A)**: Serving as the audio modality baseline, this model employs two GRU layers stacked on top of the extracted Low-Level Descriptors (LLDs). The GRU layers are tasked with capturing temporal dependencies in the audio signal, facilitating the prediction of arousal and valence based solely on auditory cues. This unimodal approach allows us to assess the contribution of the audio modality in isolation.

4 **VGGFace-RNN (V)**: This model mirrors the architecture of **AffWildNet + Dynamic (V)**, utilizing the VGGFace network for visual feature extraction followed by two GRU layers for temporal modeling. Additionally, it is trained on the Aff-Wild2 dataset, enhancing its ability to generalize across the diverse conditions present in real-world scenarios. This unimodal approach provides a benchmark for assessing the performance of the visual modality alone.

5 **VGGFace-RNN (A + V)**: Representing the audio-visual baseline, this model concatenates the outputs from the audio-only **RNN (A)** and video-only **VGGFace-RNN (V)** models. The concatenated features are then processed through two additional GRU layers, followed by a dense layer for the final predictions. This architecture facilitates the integration of audio and visual information, providing a baseline for multimodal fusion against which **ExpTrm** is compared. The concatenation approach serves as a straightforward method for multimodal integration, allowing us to evaluate the effectiveness of more sophisticated fusion techniques employed by **ExpTrm**.

6 **NISL (V)**: Deng *et al.* [48] introduce pretrained Convolutional Neural Network (CNN) and Recurrent Neural Network (RNN) models trained within a multitask framework. These models are designed to simultaneously predict emotional categories, facial action units, and emotional attributes, leveraging external datasets to balance the distribution of emotions within Aff-Wild2.

doi:10.20944/preprints202503.1265.v1

This comprehensive approach enhances the model's capacity to generalize and accurately recognize a wide array of emotional expressions by leveraging shared representations across related tasks.

7 M^3**T (A + V)**: Zhang *et al.* [49] propose a multimodal model utilizing a 3D convolutional network coupled with a bidirectional RNN. The model is trained in a multitask setting, jointly predicting emotional categories alongside valence attributes. Although an attention mechanism is incorporated atop the recurrent layers to facilitate audio-visual feature fusion, empirical results indicate that simple concatenation of features yields superior performance. Notably, Zhang *et al.* [49] observe that their attention-based fusion may not have achieved full model convergence compared to concatenation-based approaches. In contrast, our preliminary experiments demonstrate that incorporating a cross-modal attention layer within **ExpTrm** offers tangible benefits over simple feature concatenation, thereby enhancing expression detection accuracy.

It is important to highlight that both **NISL (V)** and M^3**T (A + V)** are external studies that report state-of-the-art results on the Aff-Wild2 database. Our proposed baseline, **VGGFace-RNN (A + V)**, undergoes thorough training on the Aff-Wild2 dataset, making it directly comparable to these external benchmarks. While all state-of-the-art models incorporate convolutional layers for processing video features, the primary distinction of our **ExpTrm** model lies in the utilization of dot-product attention mechanisms for multimodal fusion. Additionally, we conduct an analysis of the individual contributions of each modality within the multimodal architecture.

5. Experimental Results

5.1. Baseline Model Performance Analysis

Initially, we assess the performance of various baseline models as outlined in Section 4.5. The AffWildNet models, which are exclusively trained on the video modality, demonstrate that increasing the number of neurons in the fully connected (FC) layer from 2000 to 4096 leads to noticeable performance enhancements. Specifically, the AffWildNet + Static (V) model with an FC layer of 4096 dimensions achieves a higher Concordance Correlation Coefficient (CCC) for both valence (0.244) and arousal (0.297) compared to its 2000-dimensional counterpart (0.176 for Va and 0.198 for Ar).

Furthermore, the transition from static to dynamic models within the AffWildNet framework yields substantial improvements. The AffWildNet + Dynamic (V) model, which incorporates recurrent layers to capture temporal dependencies, registers a 7% absolute increase in valence (0.315) and a 6% absolute improvement in arousal (0.357) over the static model with 4096-dimensional FC layers. This underscores the significance of temporal context in enhancing emotion recognition accuracy.

The VGGFace-RNN (V) model, retrained on the Aff-Wild2 dataset, exhibits a remarkable 12.4% absolute improvement in arousal (0.481) and a 6.4% absolute enhancement in valence (0.379) compared to the AffWildNet + Dynamic (V) model. This substantial gain highlights the effectiveness of leveraging pre-trained networks like VGGFace for feature extraction, which provides richer and more discriminative visual representations essential for accurate emotion detection.

In contrast, the audio-only baseline, RNN (A), demonstrates inferior performance relative to the video-only models, with CCC scores of 0.105 for valence and 0.168 for arousal. This aligns with observations from previous studies [49], suggesting that visual modalities, particularly facial expressions, play a more dominant role in conveying emotional states within the Aff-Wild2 dataset. The limited performance of the audio modality may be attributed to the nature of the dataset, where expressions are predominantly conveyed through visual cues, and audio cues provide supplementary information.

When combining audio and visual modalities, the VGGFace-RNN (A+V) model shows an improvement in arousal prediction by 2.5% (CCC-Ar = 0.506) compared to the video-only model (VGGFace-RNN (V) with CCC-Ar = 0.481). However, it exhibits a slight decrease in valence prediction by 3.5% (CCC-Va = 0.344) compared to the same video-only baseline. These mixed results indicate that while the integration of audio and visual modalities can enhance certain aspects of emotion recognition, it may also introduce challenges in balancing the contributions of each modality, particularly for valence estimation.

Among external studies, the NISL (V) model [48] achieves a CCC of 0.373 for valence and 0.513 for arousal, showing slight improvements over some of the established baselines. Conversely, the M^3T (A+V) model [49] demonstrates a significantly better performance for arousal (0.550) but a lower performance for valence (0.320) compared to other models. This discrepancy suggests that while some multimodal models excel in certain emotional dimensions, their overall efficacy may vary based on the specific architectural choices and training methodologies employed.

Overall, these baseline scores provide a comprehensive benchmark against which the performance of our proposed **ExpTrm** model can be evaluated. The established baselines, encompassing both unimodal and multimodal approaches, highlight the strengths and limitations of different modalities and fusion strategies in the context of continuous emotion recognition.

5.2. Performance of the Proposed ExpTrm Model

Subsequently, we evaluate the performance of the proposed **ExpTrm** model in predicting valence and arousal values, as detailed in Table 1. The results indicate that **ExpTrm** significantly outperforms the VGGFace-RNN (V) baseline in both valence and arousal predictions. Specifically, **ExpTrm (V)** achieves a CCC of 0.381 for valence and 0.489 for arousal, marking improvements of 0.2% and 0.8% respectively over the VGGFace-RNN (V) baseline.

Table 1. Performance Evaluation on the Aff-Wild2 Database. Significance analysis was conducted using Fisher's z-transformation ($p < 0.01$). **Ar** denotes Arousal and **Va** denotes Valence. † Indicates significant differences compared to corresponding VGGFace-RNN models.

Model	**CCC-Va**	**CCC-Ar**
AffWildNet + Static (V) (2000) [47]	0.176	0.198
AffWildNet + Static (V) (4096) [47]	0.244	0.297
AffWildNet + Dynamic (V)	0.315	0.357
RNN (A)	0.105	0.168
VGGFace-RNN (V)	0.379	0.481
VGGFace-RNN (A+V)	0.344	0.506
NISL (V) [48]	0.373	0.513
M^3T (A+V) [49]	0.320	0.550
ExpTrm (V)	**0.381**	**0.489**†
ExpTrm (A+V)	**0.396**†	**0.525**†

When incorporating both audio and visual modalities, **ExpTrm (A+V)** further enhances performance, attaining CCC scores of 0.396 for valence and 0.525 for arousal. These represent absolute gains of 1.5% and 3.6% over the **ExpTrm (V)** model, underscoring the advantages of cross-modal attention mechanisms in effectively integrating complementary information from audio and visual streams.

Overall, the **ExpTrm (A+V)** model achieves absolute gains of 1.7% in valence and 1.9% in arousal compared to the VGGFace-RNN (A+V) baseline. These improvements demonstrate the efficacy of the transformer-based architecture in capturing complex emotional cues through sophisticated multimodal feature integration. Additionally, the **ExpTrm** models exhibit competitive performance relative to external state-of-the-art models, particularly in arousal prediction, where **ExpTrm (A+V)** closely approaches the performance of the M^3T (A+V) model.

It is noteworthy that the proposed **ExpTrm** models and the baseline VGGFace-RNN models possess a similar number of parameters, with discrepancies of less than 10%. This parity ensures that the observed performance enhancements are attributable to the architectural innovations of **ExpTrm**, specifically the cross-modal attention mechanisms, rather than merely increased model capacity.

Compared to external studies, our **ExpTrm** models offer competitive results, particularly for valence prediction. While some external models like M^3T excel in arousal prediction, our **ExpTrm** models demonstrate a balanced improvement across both emotional dimensions. It is important to acknowledge that external models often incorporate additional data and employ multi-task learning frameworks, which could further enhance their performance. Nonetheless, the promising results of **ExpTrm** suggest that integrating cross-modal attention within transformer architectures holds significant potential for advancing multimodal emotion recognition.

5.3. Ablation Studies

To comprehensively understand the contributions of each modality and the effectiveness of the cross-modal attention mechanisms within the **ExpTrm** architecture, we conduct an extensive ablation study. This study simulates real-world scenarios where one of the modalities might be missing or occluded, such as when a user moves out of the camera's field of view or when audio input is unreliable or absent.

The ablation study involves systematically masking the inputs from either the audio or visual modality by replacing them with zero tensors. The proportion of missing data is varied incrementally from 0% (no data missing) to 100% (complete absence of the modality) to evaluate the model's performance under varying degrees of modality loss. Each masking scenario is randomly applied, and the experiments are repeated across 10 trials on the test set to ensure statistical robustness and to account for variability in data distribution.

Impact of Modality Absence

The results of the ablation study reveal critical insights into the dependency of the **ExpTrm** model on each modality:

- **Absence of Visual Modality (Audio-Only)**: When the visual modality is entirely absent (100% masking), the **ExpTrm** model's performance degrades significantly, with CCC values for valence and arousal approaching zero. This substantial decline indicates that, within the Aff-Wild2 dataset, visual cues are paramount for accurate emotion recognition. The absence of facial expressions severely hampers the model's ability to infer emotional states based solely on audio, highlighting the limited expressiveness of audio cues in this context.
- **Partial Absence of Visual Modality**: As the proportion of missing visual data increases, there is a corresponding linear decline in performance. This trend underscores the model's reliance on visual information for maintaining prediction accuracy. However, even with partial visual data loss, the model retains a moderate level of performance, demonstrating some resilience to incomplete visual inputs.
- **Absence of Audio Modality (Visual-Only)**: In contrast, the complete absence of the audio modality results in only a marginal decrease in performance, approximately 2%, for both valence and arousal predictions. This slight decline suggests that while audio cues contribute to emotion recognition, the visual modality carries the bulk of the informative signals. The model's ability to maintain near-baseline performance in the absence of audio indicates that visual features are sufficiently robust for accurate emotion detection in most cases.
- **Partial Absence of Audio Modality**: When the audio data is partially missing, the model exhibits a gradual decrease in performance, albeit much less pronounced compared to the loss of visual data. This observation highlights the model's capacity to leverage the remaining audio information to supplement the visual cues, thereby mitigating the impact of incomplete audio inputs.

Robustness and Adaptability

The ablation study conclusively demonstrates that the **ExpTrm** architecture is highly dependent on the visual modality for optimal performance in emotion recognition tasks. The substantial performance degradation in the absence of visual data underscores the necessity of rich visual information, particularly facial expressions, in accurately inferring emotional states within the Aff-Wild2 dataset. However, the model's resilience in scenarios where audio data is missing or partially absent reflects its adaptability and the effective integration of multimodal information through cross-modal attention mechanisms.

These findings highlight the importance of visual cues in emotion recognition and suggest that while audio cues can enhance performance, especially in certain emotional dimensions like arousal, they are not as critical as visual information in the given dataset. Consequently, future enhancements to the **ExpTrm** model could explore strategies to bolster audio feature utilization or integrate additional modalities to further improve robustness and performance across diverse scenarios.

Overall Implications

The ablation results emphasize the critical role of visual information in the **ExpTrm** model's success. While the model demonstrates some level of flexibility in handling missing audio data, the dependency on visual cues suggests that enhancements in visual feature extraction or the incorporation of more sophisticated visual processing techniques could further elevate the model's performance. Additionally, exploring techniques to more effectively leverage audio cues when available could help in achieving a more balanced and comprehensive emotion recognition system.

5.4. Statistical Significance and Comparative Analysis

To ensure the validity and reliability of the observed performance improvements, statistical significance analyses were conducted using Fisher's z-transformation. The results, as presented in Table 1, indicate that the performance gains achieved by the **ExpTrm** models are statistically significant compared to the corresponding VGGFace-RNN baselines, particularly for arousal predictions. The **ExpTrm (V)** model shows a significant improvement in arousal prediction (CCC-Ar = 0.489) compared to the VGGFace-RNN (V) baseline (CCC-Ar = 0.481), and the multimodal **ExpTrm (A+V)** model exhibits significant gains in both valence (CCC-Va = 0.396) and arousal (CCC-Ar = 0.525).

These significant improvements affirm that the architectural innovations introduced in **ExpTrm**, specifically the cross-modal attention mechanisms, effectively enhance the model's capacity to integrate and prioritize multimodal information. The statistically significant gains validate the superiority of the **ExpTrm** model over traditional RNN-based approaches and underscore its potential as a robust solution for continuous emotion recognition in real-world applications.

5.5. Discussion

The experimental results unequivocally demonstrate that the proposed **ExpTrm** model outperforms existing baseline and state-of-the-art models in the task of continuous emotion recognition on the Aff-Wild2 dataset. The dual advantage of leveraging both audio and visual modalities through sophisticated cross-modal attention mechanisms enables **ExpTrm** to capture a more nuanced and comprehensive representation of user expressions.

Valence and Arousal Predictions

Valence and arousal are two fundamental dimensions in emotion representation, capturing the positivity/negativity and the intensity of emotions, respectively. The **ExpTrm** model's superior performance in both dimensions indicates its effectiveness in discerning subtle emotional variations. The marginal improvement in valence prediction when integrating audio cues suggests that while audio information provides additional context, the primary driver for valence estimation remains the visual modality. Conversely, the more substantial improvement in arousal prediction underscores the complementary nature of audio features in capturing the intensity of emotions, which is often reflected in vocal tone and speech patterns.

Multimodal Integration Benefits

The enhanced performance of the multimodal **ExpTrm (A+V)** model compared to its unimodal counterparts highlights the synergistic benefits of integrating audio and visual information. Cross-modal attention allows the model to dynamically allocate focus to the most informative features from each modality, thereby improving the overall accuracy and robustness of emotion recognition. This integration is particularly beneficial in scenarios where one modality may provide ambiguous or incomplete information, as the complementary modality can compensate, leading to more reliable predictions.

Comparison with External State-of-the-Art Models

When juxtaposed with external state-of-the-art models such as NISL (V) [48] and M^3T (A+V) [49], **ExpTrm** holds its ground by delivering competitive results, especially in arousal prediction. While some models like M^3T exhibit superior performance in specific dimensions, **ExpTrm** offers a balanced improvement across both valence and arousal, making it a versatile model for comprehensive emotion recognition tasks. Furthermore, the ability of **ExpTrm** to achieve these results without relying on additional data or multi-task learning frameworks positions it as a promising candidate for practical applications where computational resources and data availability may be constrained.

Model Complexity and Efficiency

Despite the substantial performance gains, it is noteworthy that the **ExpTrm** models maintain a similar number of parameters compared to the baseline VGGFace-RNN models, with less than a 10% difference. This efficiency is attributed to the optimized design of the transformer architecture, which effectively captures complex dependencies without excessively increasing model size. The use of pre-trained networks for feature extraction further contributes to this efficiency by providing rich representations that enhance the model's learning capacity without necessitating a proportional increase in parameters.

6. Conclusion and Future Directions

In this study, we introduced **ExpTrm**, a novel multimodal architecture designed for robust expression detection using the Aff-Wild2 database. Our approach leverages the transformer architecture, enhanced with a cross-modal attention layer that effectively identifies and emphasizes pertinent cues from both audio and visual modalities. By implementing a straightforward late fusion strategy, **ExpTrm** integrates these cues to accurately predict arousal and valence scores for each video frame within the dataset.

The experimental evaluations conducted demonstrate that **ExpTrm** significantly outperforms competitive baseline models and previously established methods, achieving state-of-the-art performance in the task of expression detection. This superior performance underscores the efficacy of incorporating cross-modal attention mechanisms in transformer-based architectures, enabling the model to capture and integrate complex emotional signals from both audio and visual inputs. Furthermore, our analysis revealed that the visual modality contributes more substantially to the model's performance, providing rich and detailed cues that enhance the accuracy of emotion recognition.

The implications of our findings are far-reaching, suggesting that **ExpTrm** can be effectively deployed in real-world applications where understanding and responding to human emotions is crucial. For instance, in interactive systems such as virtual assistants, healthcare monitoring tools, and educational platforms, the ability to accurately detect and interpret user emotions can lead to more personalized and empathetic user experiences.

Despite the promising results, our study acknowledges certain limitations. The heavy reliance on visual cues indicates that **ExpTrm** may face challenges in environments where visual data is compromised, such as in low-light conditions or when subjects are partially obscured. Additionally, while the integration of audio and visual modalities has proven beneficial, the current model does not exploit other potentially informative modalities, such as physiological signals or contextual information from the surrounding environment.

Building on the successes of **ExpTrm**, our future work aims to explore several avenues to further enhance the model's capabilities and applicability:

1. **Extension to Expression Classification**: While the current study focuses on continuous valence-arousal estimation, we plan to extend **ExpTrm** to handle categorical expression classification. This extension will enable the model to recognize discrete emotional states such as happiness, sadness, anger, and surprise, providing a more comprehensive understanding of user emotions.
2. **Incorporation of Facial Landmarks**: To harness additional visual information, we intend to integrate facial landmarks as supplementary inputs. By incorporating precise facial feature points, **ExpTrm** can achieve a finer-grained analysis of facial expressions, potentially improving the detection of subtle emotional nuances.
3. **Robust Training with Missing Data**: Recognizing that real-world scenarios often involve incomplete or missing data, we will investigate various training strategies to enhance **ExpTrm**'s robustness. Techniques such as data augmentation, imputation methods, and specialized loss functions will be explored to ensure the model maintains high performance even when one of the modalities is partially or entirely unavailable.
4. **Integration of Additional Modalities**: To further enrich the model's emotional understanding, we plan to incorporate additional modalities such as textual data and physiological signals. Integrating text from speech transcripts or contextual information can provide deeper insights into the user's emotional state, while physiological data like heart rate and skin conductance can offer objective measures of emotional arousal.
5. **Real-Time Implementation and Optimization**: To facilitate the deployment of **ExpTrm** in interactive systems, we aim to optimize the model for real-time processing. This will involve streamlining the architecture, reducing computational overhead, and ensuring efficient memory usage without compromising accuracy.
6. **Personalization and Adaptability**: Emotions are inherently personal and can vary significantly across individuals. Future research will focus on developing personalized models that adapt to individual differences in emotional expression. Techniques such as transfer learning and adaptive algorithms will be employed to tailor **ExpTrm** to specific users, enhancing its accuracy and relevance.
7. **Enhanced Multimodal Fusion Techniques**: While cross-modal attention has proven effective, we plan to explore more sophisticated fusion strategies that can dynamically adjust the integration of modalities based on contextual factors. Approaches such as hierarchical attention mechanisms and gated fusion networks will be investigated to further improve the synergy between audio and visual inputs.
8. **Comprehensive Evaluation on Diverse Datasets**: To validate the generalizability of **ExpTrm**, we will conduct evaluations on a variety of datasets encompassing different cultures, languages, and recording conditions. This will ensure that the model performs consistently across diverse populations and real-world scenarios.

In summary, the development and evaluation of **ExpTrm** mark a significant advancement in the field of multimodal emotion recognition. By effectively integrating audio and visual modalities through cross-modal attention mechanisms, **ExpTrm** achieves high accuracy in continuous valence-arousal estimation, setting a new benchmark for future research. The proposed future directions aim to address current limitations and expand the model's capabilities, paving the way for more versatile and robust emotion detection systems that can be seamlessly integrated into a wide range of applications.

References

1. N. Cummins, S. Scherer, J. Krajewski, S. Schnieder, J. Epps, and T.F. Quatieri, "A review of depression and suicide risk assessment using speech analysis," *Speech Communication*, vol. 71, pp. 10–49, July 2015.
2. C. Clavel, I. Vasilescu, L. Devillers, G. Richard, and T. Ehrette, "Fear-type emotion recognition for future audio-based surveillance systems," *Speech Communication*, vol. 50, no. 6, pp. 487–503, June 2008.
3. D. Litman and K. Forbes-Riley, "Predicting student emotions in computer-human tutoring dialogues," in *ACM Association for Computational Linguistics (ACL 2004)*, Barcelona, Spain, July 2004, pp. 1–8.

4. J.A. Russell, "A circumplex model of affect," *Journal of Personality and Social Psychology*, vol. 39, no. 6, pp. 1161–1178, December 1980.
5. Moataz El Ayadi, Mohamed S Kamel, and Fakhri Karray, "Survey on speech emotion recognition: Features, classification schemes, and databases," *Pattern Recognition*, vol. 44, no. 3, pp. 572–587, 2011.
6. R. Cowie and R.R. Cornelius, "Describing the emotional states that are expressed in speech," *Speech Communication*, vol. 40, no. 1-2, pp. 5–32, April 2003.
7. Evangelos Sariyanidi, Hatice Gunes, and Andrea Cavallaro, "Automatic analysis of facial affect: A survey of registration, representation, and recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 37, no. 6, pp. 1113–1133, 2015.
8. Dimitrios Kollias, Mihalis A Nicolaou, Irene Kotsia, Guoying Zhao, and Stefanos Zafeiriou, "Recognition of affect in the wild using deep neural networks," in *Computer Vision and Pattern Recognition Workshops (CVPRW), 2017 IEEE Conference on*. IEEE, 2017, pp. 1972–1979.
9. H. Gunes and M. Piccardi, "A bimodal face and body gesture database for automatic analysis of human nonverbal affective behavior," in *18th International Conference on Pattern Recognition (ICPR 2006)*, Hong Kong, China, August 2006, vol. 1, pp. 1148–1153.
10. M. Coulson, "Attributing emotion to static body postures: Recognition accuracy, confusions, and viewpoint dependence," *Journal of Nonverbal Behavior*, vol. 28, no. 2, pp. 117–139, June 2004.
11. H.P. Martinez, Y. Bengio, and G.N. Yannakakis, "Learning deep physiological models of affect," *IEEE Computational Intelligence Magazine*, vol. 8, no. 2, pp. 20–33, May 2013.
12. Kyung Hwan Kim, Seok Won Bang, and Sang Ryong Kim, "Emotion recognition system using short-term monitoring of physiological signals," *Medical and biological engineering and computing*, vol. 42, no. 3, pp. 419–427, 2004.
13. A. Mehrabian, "Communication without words," in *Communication Theory*, C.D. Mortensen, Ed., pp. 193–200. Transaction Publishers, New Brunswick, NJ, USA, December 2007.
14. S. Parthasarathy and C. Busso, "Jointly predicting arousal, valence and dominance with multi-task learning," in *Interspeech 2017*, Stockholm, Sweden, August 2017, pp. 1103–1107.
15. Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, 2017, pp. 5998–6008.
16. Dimitrios Kollias and Stefanos Zafeiriou, "Aff-wild2: Extending the aff-wild database for affect recognition," *arXiv preprint arXiv:1811.07770*, 2018.
17. J. Lee, M. Reyes, T. Smyser, Y. Liang, and K. Thornburg, "SAfety VEhicles using adaptive interface technology (task 5) final report: Phase 1," Technical report, The University of Iowa, Iowa City, IA, USA, November 2004.
18. J.R.J. Fontaine, K.R. Scherer, E.B. Roesch, and P.C. Ellsworth, "The world of emotions is not two-dimensional," *Psychological Science*, vol. 18, no. 12, pp. 1050–1057, December 2007.
19. Z. Aldeneh and E. Mower Provost, "Using regional saliency for speech emotion recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2017)*, New Orleans, LA, USA, March 2017, pp. 2741–2745.
20. B. Schuller, S. Steidl, and A. Batliner, "The INTERSPEECH 2009 emotion challenge," in *Interspeech 2009 - Eurospeech*, Brighton, UK, September 2009, pp. 312–315.
21. B. Schuller, S. Steidl, A. Batliner, F. Burkhardt, L. Devillers, C. Muller, and S. Narayanan, "The INTERSPEECH 2010 paralinguistic challenge," in *Interspeech 2010*, Makuhari, Japan, September 2010, pp. 2794–2797.
22. Z. Zeng, M. Pantic, G.I. Roisman, and T.S. Huang, "A survey of affect recognition methods: Audio, visual, and spontaneous expressions," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 1, pp. 39–58, January 2009.
23. Caifeng Shan, Shaogang Gong, and Peter W McOwan, "Facial expression recognition based on local binary patterns: A comprehensive study," *Image and vision Computing*, vol. 27, no. 6, pp. 803–816, 2009.
24. Ping Liu, Shizhong Han, Zibo Meng, and Yan Tong, "Facial expression recognition via a boosted deep belief network," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 1805–1812.
25. Heechul Jung, Sihaeng Lee, Junho Yim, Sunjeong Park, and Junmo Kim, "Joint fine-tuning in deep neural networks for facial expression recognition," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 2983–2991.
26. Ya Chang, Changbo Hu, Rogerio Feris, and Matthew Turk, "Manifold based analysis of facial expression," *Image and Vision Computing*, vol. 24, no. 6, pp. 605–614, 2006.

27. Maja Pantic and Ioannis Patras, "Dynamics of facial expression: recognition of facial actions and their temporal segments from face profile image sequences," *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 36, no. 2, pp. 433–449, 2006.
28. Ali Mollahosseini, David Chan, and Mohammad H Mahoor, "Going deeper in facial expression recognition using deep neural networks," in *2016 IEEE Winter conference on applications of computer vision (WACV)*. IEEE, 2016, pp. 1–10.
29. Ali Mollahosseini, Behzad Hasani, and Mohammad H Mahoor, "Affectnet: A database for facial expression, valence, and arousal computing in the wild," *IEEE Transactions on Affective Computing*, vol. 10, no. 1, pp. 18–31, 2017.
30. Samuel Albanie, Arsha Nagrani, Andrea Vedaldi, and Andrew Zisserman, "Emotion recognition in speech using cross-modal transfer in the wild," in *Proceedings of the 26th ACM international conference on Multimedia*, 2018, pp. 292–301.
31. Xi Ouyang, Shigenori Kawaai, Ester Gue Hua Goh, Shengmei Shen, Wan Ding, Huaiping Ming, and Dong-Yan Huang, "Audio-visual emotion recognition using deep transfer learning and multiple temporal models," in *Proceedings of the 19th ACM International Conference on Multimodal Interaction*, 2017, pp. 577–582.
32. Valentin Vielzeuf, Stéphane Pateux, and Frédéric Jurie, "Temporal multimodal fusion for video emotion classification in the wild," in *Proceedings of the 19th ACM International Conference on Multimodal Interaction*, 2017, pp. 569–576.
33. Panagiotis Tzirakis, George Trigeorgis, Mihalis A Nicolaou, Björn W Schuller, and Stefanos Zafeiriou, "End-to-end multimodal emotion recognition using deep neural networks," *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 8, pp. 1301–1309, 2017.
34. Mingyi Chen, Xuanji He, Jing Yang, and Han Zhang, "3-d convolutional recurrent neural networks with attention model for speech emotion recognition," *IEEE Signal Processing Letters*, vol. 25, no. 10, pp. 1440–1444, 2018.
35. Wang Xiaohua, Peng Muzi, Pan Lijuan, Hu Min, Jin Chunhua, and Ren Fuji, "Two-level attention with two-stage multi-task learning for facial emotion recognition," *Journal of Visual Communication and Image Representation*, vol. 62, pp. 217–225, 2019.
36. Devamanyu Hazarika, Soujanya Poria, Amir Zadeh, Erik Cambria, Louis-Philippe Morency, and Roger Zimmermann, "Conversational memory network for emotion recognition in dyadic dialogue videos," in *Proceedings of the conference. Association for Computational Linguistics. North American Chapter. Meeting*. NIH Public Access, 2018, vol. 2018, p. 2122.
37. Seyedmahdad Mirsamadi, Emad Barsoum, and Cha Zhang, "Automatic speech emotion recognition using recurrent neural networks with local attention," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 2227–2231.
38. Yuanchao Li, Tianyu Zhao, and Tatsuya Kawahara, "Improved end-to-end speech emotion recognition using self attention mechanism and multitask learning.," 2019.
39. Wasifur Rahman, Md Kamrul Hasan, Sangwu Lee, AmirAli Bagher Zadeh, Chengfeng Mao, Louis-Philippe Morency, and Ehsan Hoque, "Integrating multimodal information in large pretrained transformers," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Online, July 2020, pp. 2359–2369, Association for Computational Linguistics.
40. Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.
41. Georgios Paraskevopoulos, Srinivas Parthasarathy, Aparna Khare, and Shiva Sundaram, "Multimodal and multiresolution speech recognition with transformers," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Online, July 2020, pp. 2381–2387, Association for Computational Linguistics.
42. Srinivas Parthasarathy and Shiva Sundaram, "Training strategies to handle missing modalities for audio-visual expression recognition," *arXiv preprint arXiv:2010.00734*, 2020.
43. Dimitrios Kollias and Stefanos Zafeiriou, "Expression, affect, action unit recognition: Aff-wild2, multi-task learning and arcface," *arXiv preprint arXiv:1910.04855*, 2019.
44. Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C. Berg, "Ssd: Single shot multibox detector," *Lecture Notes in Computer Science*, p. 21–37, 2016.
45. B. Schuller, S. Steidl, A. Batliner, A. Vinciarelli, K. Scherer, F. Ringeval, M. Chetouani, F. Weninger, F. Eyben, E. Marchi, M. Mortillaro, H. Salamin, A. Polychroniou, F. Valente, and S. Kim, "The INTERSPEECH 2013

computational paralinguistics challenge: Social signals, conflict, emotion, autism," in *Interspeech 2013*, Lyon, France, August 2013, pp. 148–152.

46. Omkar M. Parkhi, Andrea Vedaldi, and Andrew Zisserman, "Deep face recognition," in *British Machine Vision Conference*, 2015.
47. Dimitrios Kollias, Panagiotis Tzirakis, Mihalis A Nicolaou, Athanasios Papaioannou, Guoying Zhao, Björn Schuller, Irene Kotsia, and Stefanos Zafeiriou, "Deep affect prediction in-the-wild: Aff-wild database and challenge, deep architectures, and beyond," *International Journal of Computer Vision*, pp. 1–23, 2019.
48. Didan Deng, Zhaokang Chen, and Bertram E. Shi, "Multitask emotion recognition with incomplete labels," 2020.
49. Yuan-Hang Zhang, Rulin Huang, Jiabei Zeng, Shiguang Shan, and Xilin Chen, "m^3t: Multi-modal continuous valence-arousal estimation in the wild," 2020.
50. K. Sridhar, S. Parthasarathy, and C. Busso, "Role of regularization in the prediction of valence from speech," in *Interspeech 2018*, Hyderabad, India, September 2018, pp. 941–945.
51. 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.
52. 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.
53. 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.
54. 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).
55. 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.
56. 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.
57. 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)*.
58. 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.
59. Ilya Loshchilov and Frank Hutter. 2019. Decoupled Weight Decay Regularization. In *International Conference on Learning Representations*.
60. 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).
61. Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, may 2015. https://doi.org/10.1038/nature14539. URL http://dx.doi.org/10.1038/nature14539.
62. 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/.
63. 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.
64. 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.

65. J Ngiam, A Khosla, and M Kim. Multimodal Deep Learning. In *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, pp. 689—-696, 2011. URL http://ai.stanford.edu/{~}ang/papers/icml11-MultimodalDeepLearning.pdf.
66. 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. https://doi.org/10.1109/IJCNN.2013.6706748. URL http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6706748.
67. Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.
68. Richard Socher, Milind Ganjoo, Christopher D Manning, and Andrew Ng. Zero-Shot Learning Through Cross-Modal Transfer. In C J C Burges, L Bottou, M Welling, Z Ghahramani, and K Q Weinberger (eds.), *Advances in Neural Information Processing Systems 26*, pp. 935–943. Curran Associates, Inc., 2013. URL http://papers.nips.cc/paper/5027-zero-shot-learning-through-cross-modal-transfer.pdf.
69. 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.
70. 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.
71. 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.
72. 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.
73. 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.
74. 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.
75. Ling Zhuang, Hao Fei, and Po Hu. Knowledge-enhanced event relation extraction via event ontology prompt. *Inf. Fusion*, 100:101919, 2023.
76. 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.
77. 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.
78. 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.
79. 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.
80. 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.
81. 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.
82. 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.
83. 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.
84. Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*, 2016.

85. 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.
86. 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.
87. 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.
88. 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. https://doi.org/10.18653/v1/N19-1423. URL https://aclanthology.org/N19-1423.
89. Shengqiong Wu, Hao Fei, Leigang Qu, Wei Ji, and Tat-Seng Chua. Next-gpt: Any-to-any multimodal llm. *CoRR*, abs/2309.05519, 2023.
90. 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.
91. 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.
92. Naman Jain, Pranjali Jain, Pratik Kayal, Jayakrishna Sahit, Soham Pachpande, Jayesh Choudhari, et al. Agribot: agriculture-specific question answer system. *IndiaRxiv*, 2019.
93. 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.
94. 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.
95. 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.
96. 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.
97. 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.
98. 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.
99. 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.
100. Sanjeev Arora, Yingyu Liang, and Tengyu Ma. A simple but tough-to-beat baseline for sentence embeddings. In *ICLR*, 2017.
101. 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.
102. 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.
103. 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.
104. 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.

doi:10.20944/preprints202503.1265.v1

105. 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.
106. 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.
107. 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.
108. 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.
109. 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.
110. 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.
111. 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.
112. 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.
113. 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.
114. 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.
115. 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.
116. Bart Van Merriënboer, Dzmitry Bahdanau, Vincent Dumoulin, Dmitriy Serdyuk, David Warde-Farley, Jan Chorowski, and Yoshua Bengio. Blocks and fuel: Frameworks for deep learning. *arXiv preprint arXiv:1506.00619*, 2015.
117. Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3156–3164, 2015.
118. 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.

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