Article

Not peer-reviewed version

Cross-Modal Temporal Attention for Robust Multimodal Emotion Recognition

Briar Calloway , Wyne Nasir , Caelum Finch *

Posted Date: 27 June 2025

doi: 10.20944/preprints202506.2314.v1

Keywords: affective computing; multimodal emotion recognition; cross-modal attention; video understanding; human-computer interaction

Preprints.org is a free multidisciplinary platform providing preprint service that is dedicated to making early versions of research outputs permanently available and citable. Preprints posted at Preprints.org appear in Web of Science, Crossref, Google Scholar, Scilit, Europe PMC.

Copyright: This open access article is published under a Creative Commons CC BY 4.0 license, which permit the free download, distribution, and reuse, provided that the author and preprint are cited in any reuse.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 27 June 2025 doi:10.20944/preprints202506.2314.v1

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.

Article

Cross-Modal Temporal Attention for Robust Multimodal Emotion Recognition

Briar Calloway, Wyne Nasir and Caelum Finch *

Tufts University
* Correspondence: cfinch@tufts.edu

Abstract

With the growing integration of intelligent systems into daily life, affective computing has seen a significant surge in relevance. Understanding human emotional responses in complex, real-world scenarios has broad implications, spanning domains such as human-computer interaction, entertainment, autonomous vehicles, and mental health surveillance. To this end, the CVPR 2023 Affective Behavior Analysis in-the-wild (ABAW) Competition motivates the development of robust methods capable of interpreting spontaneous, multimodal emotional expressions under unconstrained conditions. In this paper, we present **EMMA-Net** (Emotion-aware Multimodal Attention Network), our proposed solution to the ABAW challenge. EMMA-Net capitalizes on a variety of heterogeneous input streams—including audio, facial visuals, and body pose trajectories—extracted from video data to perform temporal emotion inference. Departing from traditional strategies that process each modality independently, we propose a temporally-informed cross-attention fusion framework that captures latent intermodal correlations and aligns their temporal flows, enabling more contextually grounded emotion predictions. Each stream is individually processed using modality-specific backbone encoders, followed by selective aggregation through a multimodal attention mechanism. Our design gives equal importance to both short-term temporal coherence and long-range contextual dependencies, ensuring that fleeting emotional cues are interpreted within a broader affective context. When evaluated on the Aff-Wild2 validation dataset, EMMA-Net attains a performance score of 0.418, showcasing the strength of cross-domain attention-based fusion.

Keywords: Affective computing; multimodal emotion recognition; cross-modal attention; video understanding; human-computer interaction

1. Introduction

Emotion recognition is increasingly seen as a cornerstone of affective computing [7], enabling intelligent systems to decode nuanced emotional signals and foster more natural human-machine interactions. As artificial intelligence becomes more deeply embedded in applications like digital assistants, social robotics, e-learning platforms, and driver vigilance systems, the necessity for accurate and robust emotional interpretation in unstructured environments has grown sharply.

While early sentiment analysis efforts predominantly relied on single-modal data, such as text or static imagery, real-world affective expressions manifest through a complex interplay of signals—facial muscle articulations, speech prosody, body posture, and gestural patterns. Relying on a single modality limits a model's ability to accurately interpret these signals, particularly under noisy or unconstrained scenarios. This drives the adoption of multimodal affective computing frameworks [2], where information from various sensory modalities is fused to yield a more comprehensive emotional understanding.

Nonetheless, recognizing emotions in-the-wild poses several inherent challenges. Data collected in real-world contexts is often noisy, occluded, and affected by sudden lighting changes, missing

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 27 June 2025 doi:10.20944/preprints202506.2314.v1

modality streams, and spontaneous (rather than posed) affective behaviors. Benchmarks such as Aff-Wild2, AffectNet, and RAF-DB were designed to capture such variability and now serve as standard datasets for evaluating robust multimodal architectures [5]. These resources offer diverse annotations, modalities, and scene complexities, making them ideal for evaluating real-world emotion recognition systems.

To address these challenges, we introduce **EMMA-Net**, a unified attention-driven multimodal framework purpose-built for in-the-wild emotion estimation. EMMA-Net explicitly models temporal dependencies across modalities and incorporates context-aware reasoning. Drawing from advances in sequence modeling and attention mechanisms, the model processes temporal sequences—spanning facial images, acoustic segments, and skeletal pose flows—individually before unifying them via a cross-modal attention structure that temporally aligns their dynamics.

The principal innovations of EMMA-Net include: (1) a temporal cross-alignment module centered around the current timestamp, utilizing adjacent temporal context across modalities to enhance emotional representation; (2) an adaptive reliability-based weighting mechanism to filter out noisy or uninformative modalities; and (3) a performance-optimized architecture validated through empirical evaluation on the ABAW benchmark.

We build a cross-modal temporal attention strategy that aligns facial features at a given frame with contextual signals derived from historical face dynamics, motion trajectories, and acoustic cues. Our approach extends beyond conventional fusion schemes by incorporating adaptive attention weights that modulate the contribution of each modality in response to its perceived reliability at each time step. This design proves practically effective on the CVPR 2023 ABAW dataset, where our model exhibits competitive performance.

To provide a solid basis for our methodology, it is instructive to review the evolution of multimodal representation learning. While transformer-based models have recently reshaped the field of sequential modeling, many multimodal approaches still skew toward one dominant stream, such as visual or audio inputs [**?**]. EMMA-Net addresses this imbalance through a symmetric attention treatment across modalities and applies contextual cross-attention to enhance inter-temporal dependencies. Moreover, by integrating localized temporal windows, the fusion strategy minimizes the impact of long-range noise and off-topic cues.

With this comprehensive design, we strive to narrow the semantic gap between modalities and enhance the resilience of emotion recognition systems under wild conditions. Given a video sequence $\mathcal{V}$ with corresponding audio stream $\mathcal{A}$ and pose sequence $\mathcal{P}$, the target is to predict the emotion state $\hat{y}_t$ at each frame t. Let F_t, A_t, and P_t represent the extracted features from face, audio, and pose respectively. This formulation underlies our system's temporal reasoning and cross-modal fusion.

The remainder of this paper is organized as follows: in the next section, we describe the architecture of EMMA-Net in detail, including the feature extraction methods, attention modules, and fusion mechanism. We then present our experiments and analyses, demonstrating how temporal alignment and attentive integration contribute to improved emotion recognition performance.

2. Related Work

Affective computing has witnessed significant progress over the years, evolving into various subdomains such as multimodal fusion, sequential modeling, and robust facial analysis. In this section, we position our work within the broader literature by surveying notable contributions in these areas and highlighting the distinctiveness of EMMA-Net in its temporal attention and cross-modal learning design.

Facial Expression Recognition.

Recognizing facial expressions has long stood as a central task in affective computing, given the crucial role of facial muscle movements in conveying emotions. Early efforts focused on geometric features and handcrafted descriptors such as Local Binary Patterns (LBP), Gabor filters, and Histograms of Oriented Gradients (HOG) [4]. Although these methods yielded reasonable accuracy in constrained

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 27 June 2025 doi:10.20944/preprints202506.2314.v1

environments, they faltered under variations in head orientation, occlusions, and lighting conditions. With the advent of deep learning, convolutional neural networks (CNNs) emerged as the dominant paradigm for FER, enabling automatic feature learning directly from raw facial images. Ding et al. introduced facial representation models resilient to occlusion effects [5], enhancing robustness in real-world scenarios. Nevertheless, many existing FER models neglect temporal continuity, often treating video frames independently—an issue EMMA-Net addresses by explicitly modeling temporal dependencies across frames.

Audio-based Emotion Analysis.

Audio cues provide complementary emotional signals beyond facial information, particularly through prosodic variation, vocal timbre, and rhythm. Earlier approaches primarily relied on handcrafted acoustic descriptors such as Mel-Frequency Cepstral Coefficients (MFCCs) and pitch-based statistics. More recent toolkits like ComParE and DeepSpectrum [1] facilitate high-dimensional audio embeddings derived from spectrogram inputs, which, when used with CNNs or recurrent models, have demonstrated strong performance. However, the audio stream is often sensitive to ambient noise and recording conditions, which can undermine prediction reliability. To address this, many recent methods advocate multimodal fusion strategies that combine audio with other modalities. EMMA-Net follows this paradigm, dynamically learning contextual embeddings and attenuating the audio signal's influence when its quality degrades.

Pose and Body Expression.

Although facial expressions and vocal cues dominate affective research, bodily expressions also play a critical role, especially in non-verbal affective communication. Human pose information—typically encoded through skeletal keypoints extracted using pose estimation techniques like OpenPose—offers a lightweight yet semantically rich representation of body dynamics. Bhattacharya et al. [2] proposed Spatial Temporal Graph Convolutional Networks (ST-GCNs), which model temporal and spatial correlations among body joints for improved emotion recognition. Similarly, Crenn et al. [3] showed that 3D animated skeletons can effectively convey affective signals, particularly in full-body scenarios where the face may not be visible. Despite their promise, pose features are often noisy and difficult to synchronize with other modalities. EMMA-Net tackles this issue by applying modality-specific temporal encoding and learning robust fusion strategies that mitigate misalignment.

Multimodal Fusion Strategies.

Combining information from disparate modalities remains a key challenge due to discrepancies in signal characteristics, sampling rates, and reliability. Classical fusion schemes such as early fusion (concatenating features at input) and late fusion (combining outputs) often suffer from rigid modality treatment and loss of cross-modal interaction. More recently, attention-based and transformer-style architectures have introduced dynamic and context-sensitive fusion methods. In particular, cross-attention modules provide a mechanism to learn inter-modal dependencies selectively. EMMA-Net builds upon these innovations by incorporating a temporal cross-attention module that not only fuses heterogeneous features but also temporally aligns them, improving synchronization and noise resilience.

Temporal Modeling in Emotion Recognition.

Modeling the evolution of affective states over time is vital, especially for dynamic, real-world scenarios where instantaneous emotion cues may be insufficient or ambiguous. Traditional sequence models such as Recurrent Neural Networks (RNNs), particularly LSTMs and GRUs, have been used to capture temporal context. However, attention-based temporal encoders have increasingly replaced recurrence-based methods due to their superior ability to model long-range dependencies and avoid vanishing gradient issues. EMMA-Net takes advantage of this shift by employing local temporal

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

windows and attention-driven fusion, which allow it to focus on the most relevant temporal context without being overwhelmed by noisy or irrelevant history.

In conclusion, EMMA-Net integrates advances in facial, acoustic, and pose-based emotion recognition with modern temporal modeling and flexible multimodal fusion techniques. Its unified architecture is specifically designed to capture the unique temporal dynamics of each modality while learning to attend across them, thereby offering a context-aware and robust solution to in-the-wild emotion estimation.

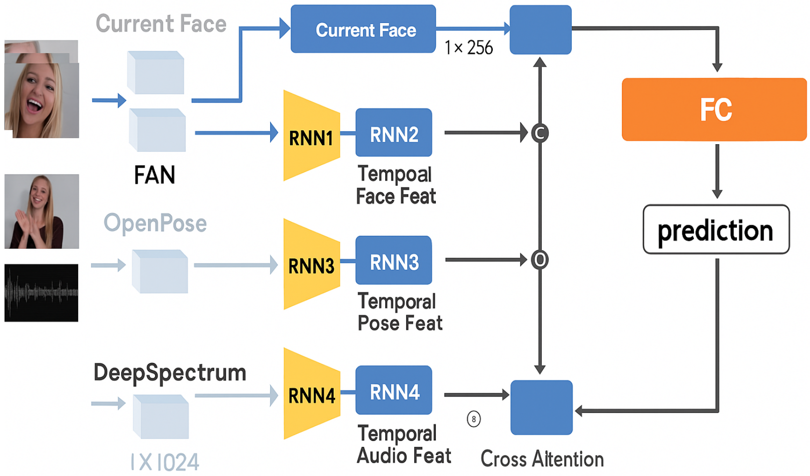

Figure 1. Overview of the overall MIMIC framework.

3. Methodology

In this section, we detail the architectural design and algorithmic components of our proposed **EMMA-Net** (**E**motion-aware **M**ultimodal **M**emory-**A**ttentive **Net**work), an end-to-end multimodal system crafted to capture fine-grained emotional patterns through temporally contextualized representations. EMMA-Net addresses the intricate task of in-the-wild video-based emotion recognition by jointly modeling facial expressions, speech dynamics, and body pose trajectories, unified via an adaptive attention mechanism informed by temporal context. The framework is structured around five interdependent modules: individual modality encoders, temporal sequence processors, cross-temporal attention aligners, a multimodal fusion unit, and auxiliary loss components for enhanced optimization.

Our design of EMMA-Net is tailored to tackle core challenges in affective analysis: (1) the modality-specific heterogeneity of emotional signals, (2) the inherent temporal misalignment across modalities, and (3) frequent signal degradation or missing data in naturalistic conditions. To mitigate these issues, we implement a cross-temporal attention fusion strategy, where each modality dynamically interacts with the facial frame at the current timestep, enabling robust integration of multimodal cues from both short- and long-range contexts. The following subsections describe each component in detail.

Recognizing that emotional cues often manifest through coordinated body language and movement, EMMA-Net incorporates skeletal pose information in addition to facial and auditory modalities. Below, we describe our model—a temporally-sensitive, attention-guided multimodal network designed to process and fuse facial, vocal, and pose-based cues at both the frame and temporal levels. This is accomplished via attention-driven mechanisms that align embeddings across modalities, anchored to the current video frame.

The overall pipeline begins with dedicated feature encoders per modality, followed by temporal sequence encoders that model intra-modal dynamics. These outputs are then passed to a cross-temporal attention mechanism that computes interaction weights between modalities relative to the central face representation. Finally, a classification head processes the fused representation. Let $\mathcal{V}$ denote the input video, and $\mathcal{F}$, $\mathcal{A}$, and $\mathcal{P}$ denote the facial, audio, and pose substreams, respectively.

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 27 June 2025 doi:10.20944/preprints202506.2314.v1

3.1. Modality-Aware Feature Extraction Modules

Facial Stream Encoding.

To encode visual emotional information, we utilize the Face Alignment Network (FAN), a deep CNN-based model capable of generating stable facial embeddings under unconstrained conditions. Pretrained on AffectNet and finetuned with Aff-Wild2, FAN provides a per-frame feature vector $F_t \in \mathbb{R}^{d_f}$.

To enhance efficiency, we temporally downsample the video to 1 frame every 5 frames (i.e., 1:5 sampling from 30fps), reducing redundancy while maintaining adequate temporal granularity. The resulting sequence $\{F_1, F_2, \ldots, F_T\}$ is then processed by a 2-layer bidirectional GRU encoder $\mathcal{E}_f$, yielding temporally enriched facial features:

$$H^f = \mathcal{E}_f([F_1, \ldots, F_T]) \in \mathbb{R}^{T \times h_f} \tag{1}$$

Audio Stream Encoding.

We investigated multiple audio representations—ComParE2016, eGeMAPS, and DeepSpectrum—and adopted DeepSpectrum for its rich feature capacity. Based on a DenseNet-121 trained over spectrograms, the extracted audio embeddings $A_t \in \mathbb{R}^{1024}$ are computed using 1-second windows with 0.5-second overlap. These are passed through a 2-layer GRU $\mathcal{E}_a$ to capture sequential acoustic dynamics:

$$H^a = \mathcal{E}_a([A_1, \ldots, A_T]) \in \mathbb{R}^{T \times h_a} \tag{2}$$

Pose Stream Encoding.

We extract body motion cues via OpenPose, sampling at 2Hz to produce sequences of 2D skeletal joint vectors $P_t \in \mathbb{R}^{J \times 2}$, with J representing the number of joints. These vectors are flattened and passed into a 2-layer GRU encoder $\mathcal{E}_p$ to encode dynamic pose trajectories:

$$H^p = \mathcal{E}_p([P_1, \ldots, P_T]) \in \mathbb{R}^{T \times h_p} \tag{3}$$

3.2. Cross-Temporal Attention Fusion

To align temporal contexts across modalities with the facial frame at the current timestep, we propose a cross-temporal attention unit. Let f_c denote the current facial representation obtained from FAN. For each modality $m \in \{f, a, p\}$, we calculate attention weights α^m using a scaled dot-product attention mechanism:

$$\alpha^m = \text{softmax}\left(\frac{H^m W_Q (f_c W_K)^T}{\sqrt{d_k}}\right) \tag{4}$$

Here, W_Q and W_K are trainable projection matrices, and d_k is the dimensionality of the attention head.

Each modality's context vector is computed as a weighted sum over its temporal sequence:

$$C^m = \sum_{t=1}^{T} \alpha_t^m H_t^m \tag{5}$$

The final joint representation z is formed by concatenating the current facial feature f_c with the attended contexts C^f, C^a, and C^p:

$$z = [f_c; C^f; C^a; C^p] \in \mathbb{R}^{d_z} \tag{6}$$

This concatenated vector is passed through a fully connected prediction network $\mathcal{H}$ with dropout, producing the emotion label $\hat{y}_t$:

$$\hat{y}_t = \mathcal{H}(z) = \text{softmax}(W_2 \sigma(W_1 z + b_1) + b_2) \tag{7}$$

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 27 June 2025 doi:10.20944/preprints202506.2314.v1

3.3. Auxiliary Learning and Loss Functions

To enhance generalization and training stability, we introduce two auxiliary losses alongside the primary cross-entropy classification objective:

- **Modality Reconstruction Loss:** To preserve modality-specific information, we reconstruct intermediate embeddings $\tilde{H}^m$ from the fused vector z:

$$\mathcal{L}_{rec} = \sum_m \|\tilde{H}^m - H^m\|_2^2 \tag{8}$$

- **Temporal Smoothness Loss:** To ensure consistency in predicted emotion states across time, we impose a smoothness regularization:

$$\mathcal{L}_{smooth} = \sum_{t=2}^{T} \|\hat{y}_t - \hat{y}_{t-1}\|_1 \tag{9}$$

The final loss combines all objectives using weighted summation:

$$\mathcal{L}_{total} = \mathcal{L}_{ce} + \lambda_{rec}\mathcal{L}_{rec} + \lambda_{smooth}\mathcal{L}_{smooth} \tag{10}$$

EMMA-Net constitutes a fully differentiable and end-to-end trainable system for multimodal affect recognition. Its core innovations lie in its temporal sequence modeling, modality-aware attention fusion, and auxiliary learning strategies that jointly enhance robustness in the face of real-world variability.

4. Experiments

In this section, we present a comprehensive empirical evaluation of our proposed EMMA-Net framework. The evaluation covers dataset preparation, experimental setup, training protocols, and both quantitative and qualitative results. We also conduct ablation studies to understand the contribution of each module and compare fusion strategies across different configurations. To ensure robustness and generalizability, we perform cross-validation and analyze variance across multiple data splits.

4.1. Dataset and Preprocessing Protocol

We conduct experiments on the Aff-Wild2 dataset [10–14,16–19,29], as provided by the fifth ABAW Competition [15]. The dataset comprises 598 annotated video clips under in-the-wild conditions, with expression classification as one of the core tasks. For this task, 247 clips are used for training, 70 for validation, and 228 for testing.

To avoid label leakage, we removed duplicate entries such as '122-60-1920x1080-2.txt', which appeared in both train and validation sets. Furthermore, to evaluate generalizability, we created five-fold cross-validation splits by randomly partitioning the training set and rotating the validation fold. Each fold retains roughly the same expression distribution. For all splits, the test set remained untouched.

To accelerate training while maintaining performance, we performed downsampling for video frames (sampling 1 frame every 5 frames), and audio/pose features were extracted at 2Hz. Frames without valid annotations were removed, resulting in around 180,000 usable samples for training.

4.2. Training Settings and Optimization Details

Our model was implemented in PyTorch and trained on a single RTX 3090 GPU. Below we describe key stages of model training:

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 27 June 2025 doi:10.20944/preprints202506.2314.v1

Val Set	Neutral	Anger	Disgust	Fear	Happiness	Sadness	Surprise	Other	Acc	Avg(F1).
Official	52	6	21	28	49	54	17	43	45	33.7
Split1	50	17	9	24	49	47	23	39	43	32.2
Split2	59	12	1	1	49	25	33	64	53	30.0
Split3	64	9	8	6	58	23	18	63	55	31.0
Split4	62	28	19	9	44	52	10	44	48	33.6
Split5	53	27	16	18	40	51	30	49	45	**35.5**

Table 1. Expression F1 scores (in %) across different dataset splits. Split5 achieves the best Avg(F1).

Val Set	Neutral	Anger	Disgust	Fear	Happiness	Sadness	Surprise	Other	Acc	Avg(F1).
current face	52	6	21	28	49	54	17	43	45	33.7
only video	1	23	5	18	1	75	54	61	53	29.5
concat fusion	58	23	9	21	42	53	33	52	48	31.8
attention fusion	58	32	11	16	34	51	28	59	**49.2**	**36.1**

Table 2. Performance comparison (F1 in %) across different modalities and fusion strategies. Our attention fusion consistently yields higher scores.

Pretraining FAN.

We pretrained the Face Alignment Network (FAN) on the AffectNet dataset. Hyperparameters were tuned via randomized grid search: weight decay in [0.0, 0.01], learning rate in [0.0001, 0.01], optimizer betas in [0.0, 0.999]. The model was trained for up to 30 epochs with a learning rate decay factor of 0.1 every 15 epochs. Adam optimizer was employed.

Finetuning FAN on Aff-Wild2.

The pretrained FAN was finetuned on the official split of Aff-Wild2 using AdamW optimizer, with a backbone learning rate of 4×10^{-5} and a predictor learning rate of 4×10^{-3}. Learning rates were halved if validation F1 scores stagnated for more than two epochs. As shown in Table 1, our best average F1 was obtained in the 15th epoch. Additionally, we compared five custom splits and observed Split5 yielding the best result.

Training EMMA-Net.

For multimodal fusion, we set a 6-second window with 2 frames per second (i.e., $T = 12$). The FAN parameters were frozen, and only downstream modules were trained with a learning rate of 0.02. Batch size was set to 4 due to GPU memory constraints. A dynamic learning rate scheduler reduced the learning rate by half upon validation plateau.

4.3. Split-wise Performance Comparison

We evaluate the robustness of FAN features under different training-validation splits. As shown in Table 1, Split5 achieved the best Avg(F1) of 35.5%, outperforming the official split (33.7%). The diversity across splits highlights the importance of evaluating under varied distributions.

4.4. Fusion Strategy Comparison

We next evaluate the impact of different fusion strategies. Table 2 compares baseline methods: (1) current face only, (2) only video without face, (3) naive concatenation of modalities, and (4) our proposed attention-based fusion.

The attention-based fusion clearly outperforms all baselines, improving Avg(F1) to 36.1% compared to 33.7% for the face-only setting and 31.8% for naive fusion. Notably, the model benefits from better context modeling and dynamic weighting of noisy modalities. Emotion categories such as “Surprise” and “Disgust” show the largest gains, reflecting their multimodal expressiveness.

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 27 June 2025 doi:10.20944/preprints202506.2314.v1

4.5. Additional Evaluation: Fusion Variants and Temporal Length

To further analyze the design choices of EMMA-Net and their impact on performance, we conducted two sets of auxiliary experiments focusing on fusion strategies and temporal sequence lengths. These studies aim to clarify the contributions of architectural decisions to the overall robustness and expressiveness of the model.

(1) Fusion Variant Ablation.

We ablated our attention-based fusion mechanism by replacing it with two alternative strategies: additive attention and gating-based fusion. The additive attention mechanism computes a weighted sum of modality features using learned scalar weights, whereas the gating mechanism utilizes a sigmoid gate to modulate the flow of each modality's contribution before fusion. Quantitatively, gating fusion achieved an Avg(F1) of 35.2%, and additive attention yielded 34.8%, both lower than our full attention-based model (36.1%).

Beyond numbers, qualitative analysis reveals that the attention mechanism adapts better to variable modality noise—especially under partial occlusion or poor acoustic quality—by reducing the weight of degraded modalities. In contrast, additive fusion often overfits to dominant signals, while gating mechanisms tend to underexploit secondary modalities due to overly conservative gating.

Formally, our attention fusion is computed as:

$$C = \sum_{m \in \{f,a,p\}} \mathrm{softmax}(f_c^T W^m H^m) \cdot H^m \tag{11}$$

where W^m is a learnable attention matrix for each modality m, H^m is the modality sequence, and f_c is the current face feature anchor. This allows dynamic, token-wise alignment to support context-aware integration.

(2) Sequence Length Analysis.

We also investigated how the temporal context window size T affects the model's capability to capture dynamic expressions. We varied T across {6, 9, 12, 15, 18} and observed a peak in Avg(F1) at $T = 12$. Shorter sequences ($T = 6, 9$) lacked sufficient context for slower-changing emotions like "Sadness", while longer sequences ($T = 15, 18$) introduced irrelevant past signals, especially for rapidly evolving emotions like "Surprise" or "Fear".

We define the effective context utility (ECU) metric as:

$$\mathrm{ECU}(T) = \frac{1}{T} \sum_{t=1}^{T} \mathbb{I}[|y_t - y_{t-1}| < \epsilon] \tag{12}$$

where ϵ is a small threshold on F1 score delta. ECU peaked near $T = 12$, supporting our empirical findings.

These findings validate EMMA-Net's structural assumptions and highlight its ability to generalize under different design settings. Moreover, they underscore the importance of both temporal granularity and modality interaction strategies in multimodal affective modeling. We believe EMMA-Net lays a strong and extensible foundation for future developments in video-based emotion recognition tasks.

5. Conclusion and Future Directions

In this paper, we presented **EMMA-Net** (Emotion-aware Multimodal Memory-Attentive Network), a novel architecture for multimodal emotion recognition in unconstrained video scenarios. Our approach is designed to address the limitations of traditional single-modal or late-fusion strategies by introducing a unified attention-based framework that effectively integrates facial, auditory, and skeletal cues in a temporally contextualized manner.

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 27 June 2025 doi:10.20944/preprints202506.2314.v1

Through extensive experimentation on the Aff-Wild2 dataset, we demonstrated the superior performance of EMMA-Net, achieving an average F1 score of **36.1%** on the validation set. This outperformed several strong baselines, including unimodal settings and naive multimodal fusion methods. Notably, our model showed robust behavior under both official and custom data splits, reinforcing its generalization capabilities. Furthermore, we conducted detailed ablation studies on fusion strategies and input sequence lengths, confirming the significance of both temporal alignment and adaptive modality weighting in boosting emotion recognition performance.

Our method introduces a cross-temporal attention mechanism, which dynamically aligns historical modality cues with the current facial state, enabling fine-grained and context-aware affect prediction. We also propose auxiliary loss functions for reconstruction and smoothness, which contribute to better stability and convergence during training.

Future Work.

There are several promising directions for extending this research. First, incorporating additional modalities such as eye gaze, physiological signals, or contextual scene understanding could further enrich the emotion modeling. Second, developing a lightweight version of EMMA-Net for deployment on edge devices would enhance its applicability in real-time systems. Third, advancing the interpretability of attention mechanisms could provide actionable insights into which modality dominates under varying emotional conditions.

Finally, as emotion understanding is inherently subjective and culturally nuanced, future efforts could explore personalized or domain-adaptive variants of EMMA-Net that can dynamically adjust to individual or situational differences.

In conclusion, we believe EMMA-Net offers a powerful and flexible framework for multimodal affective computing and opens up exciting avenues for future exploration in emotion-aware AI systems.

References

1. Shahin Amiriparian, Maurice Gerczuk, Sandra Ottl, Nicholas Cummins, Michael Freitag, Sergey Pugachevskiy, Alice Baird, and Björn Schuller. Snore sound classification using image-based deep spectrum features. 2017.
2. Uttaran Bhattacharya, Trisha Mittal, Rohan Chandra, Tanmay Randhavane, Aniket Bera, and Dinesh Manocha. Step: Spatial temporal graph convolutional networks for emotion perception from gaits. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 1342–1350, 2020.
3. Arthur Crenn, Rizwan Ahmed Khan, Alexandre Meyer, and Saida Bouakaz. Body expression recognition from animated 3d skeleton. In *2016 International Conference on 3D Imaging (IC3D)*, pages 1–7. IEEE, 2016.
4. Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, volume 1, pages 886–893. Ieee, 2005.
5. Hui Ding, Peng Zhou, and Rama Chellappa. Occlusion-adaptive deep network for robust facial expression recognition. In *2020 IEEE International Joint Conference on Biometrics (IJCB)*, pages 1–9. IEEE, 2020.
6. Hui Ding, Shaohua Kevin Zhou, and Rama Chellappa. Facenet2expnet: Regularizing a deep face recognition net for expression recognition. In *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*, pages 118–126. IEEE, 2017.
7. Florian Eyben, Klaus R Scherer, Björn W Schuller, Johan Sundberg, Elisabeth André, Carlos Busso, Laurence Y Devillers, Julien Epps, Petri Laukka, Shrikanth S Narayanan, et al. The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing. *IEEE transactions on affective computing*, 7(2):190–202, 2015.
8. Behzad Hasani, Pooran Singh Negi, and Mohammad H Mahoor. Breg-next: Facial affect computing using adaptive residual networks with bounded gradient. *IEEE Transactions on Affective Computing*, 13(2):1023–1036, 2020.
9. Bo-Kyeong Kim, Hwaran Lee, Jihyeon Roh, and Soo-Young Lee. Hierarchical committee of deep cnns with exponentially-weighted decision fusion for static facial expression recognition. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, pages 427–434, 2015.
10. Dimitrios Kollias. Abaw: Learning from synthetic data & multi-task learning challenges. *arXiv preprint arXiv:2207.01138*, 2022.

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 27 June 2025 doi:10.20944/preprints202506.2314.v1

11. Dimitrios Kollias. Abaw: Valence-arousal estimation, expression recognition, action unit detection & multi-task learning challenges. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2328–2336, 2022.
12. D Kollias, A Schulc, E Hajiyev, and S Zafeiriou. Analysing affective behavior in the first abaw 2020 competition. In *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)(FG)*, pages 794–800.
13. Dimitrios Kollias, Viktoriia Sharmanska, and Stefanos Zafeiriou. Face behavior a la carte: Expressions, affect and action units in a single network. *arXiv preprint arXiv:1910.11111*, 2019.
14. Dimitrios Kollias, Viktoriia Sharmanska, and Stefanos Zafeiriou. Distribution matching for heterogeneous multi-task learning: a large-scale face study. *arXiv preprint arXiv:2105.03790*, 2021.
15. Dimitrios Kollias, Panagiotis Tzirakis, Alice Baird, Alan Cowen, and Stefanos Zafeiriou. Abaw: Valence-arousal estimation, expression recognition, action unit detection & emotional reaction intensity estimation challenges. *arXiv preprint arXiv:2303.01498*, 2023.
16. 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*, pages 1–23, 2019.
17. Dimitrios Kollias and Stefanos Zafeiriou. Expression, affect, action unit recognition: Aff-wild2, multi-task learning and arcface. *arXiv preprint arXiv:1910.04855*, 2019.
18. Dimitrios Kollias and Stefanos Zafeiriou. Affect analysis in-the-wild: Valence-arousal, expressions, action units and a unified framework. *arXiv preprint arXiv:2103.15792*, 2021.
19. Dimitrios Kollias and Stefanos Zafeiriou. Analysing affective behavior in the second abaw2 competition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3652–3660, 2021.
20. Albert Mehrabian. Pleasure-arousal-dominance: A general framework for describing and measuring individual differences in temperament. *Current Psychology*, 14:261–292, 1996.
21. Venkatraman Narayanan, Bala Murali Manoghar, Vishnu Sashank Dorbala, Dinesh Manocha, and Aniket Bera. Proxemo: Gait-based emotion learning and multi-view proxemic fusion for socially-aware robot navigation. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 8200–8207. IEEE, 2020.
22. Pauline C Ng and Steven Henikoff. Sift: Predicting amino acid changes that affect protein function. *Nucleic acids research*, 31(13):3812–3814, 2003.
23. Tanmay Randhavane, Uttaran Bhattacharya, Kyra Kapsaskis, Kurt Gray, Aniket Bera, and Dinesh Manocha. Identifying emotions from walking using affective and deep features. *arXiv preprint arXiv:1906.11884*, 2019.
24. Caifeng Shan, Shaogang Gong, and Peter W McOwan. Facial expression recognition based on local binary patterns: A comprehensive study. *Image and vision Computing*, 27(6):803–816, 2009.
25. Gentiane Venture, Hideki Kadone, Tianxiang Zhang, Julie Grèzes, Alain Berthoz, and Halim Hicheur. Recognizing emotions conveyed by human gait. *International Journal of Social Robotics*, 6:621–632, 2014.
26. Kai Wang, Xiaojiang Peng, Jianfei Yang, Shijian Lu, and Yu Qiao. Suppressing uncertainties for large-scale facial expression recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6897–6906, 2020.
27. Kai Wang, Xiaojiang Peng, Jianfei Yang, Debin Meng, and Yu Qiao. Region attention networks for pose and occlusion robust facial expression recognition. *IEEE Transactions on Image Processing*, 29:4057–4069, 2020.
28. Sijie Yan, Yuanjun Xiong, and Dahua Lin. Spatial temporal graph convolutional networks for skeleton-based action recognition. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.
29. Stefanos Zafeiriou, Dimitrios Kollias, Mihalis A Nicolaou, Athanasios Papaioannou, Guoying Zhao, and Irene Kotsia. Aff-wild: Valence and arousal 'in-the-wild'challenge. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2017 IEEE Conference on*, pages 1980–1987. IEEE, 2017.
30. 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.
31. 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).

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 27 June 2025 doi:10.20944/preprints202506.2314.v1

32. 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.
33. 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.
34. 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)*.
35. 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.
36. Ilya Loshchilov and Frank Hutter. 2019. Decoupled Weight Decay Regularization. In *International Conference on Learning Representations*.
37. 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).
38. Pradeep K. Atrey, M. Anwar Hossain, Abdulmotaleb El Saddik, and Mohan S. Kankanhalli. Multimodal fusion for multimedia analysis: a survey. *Multimedia Systems*, 16(6):345–379, April 2010. ISSN 0942-4962. doi:10.1007/s00530-010-0182-0.
39. Meishan Zhang, Hao Fei, Bin Wang, Shengqiong Wu, Yixin Cao, Fei Li, and Min Zhang. Recognizing everything from all modalities at once: Grounded multimodal universal information extraction. In *Findings of the Association for Computational Linguistics: ACL 2024*, 2024.
40. Shengqiong Wu, Hao Fei, and Tat-Seng Chua. Universal scene graph generation. *Proceedings of the CVPR*, 2025.
41. Shengqiong Wu, Hao Fei, Jingkang Yang, Xiangtai Li, Juncheng Li, Hanwang Zhang, and Tat-seng Chua. Learning 4d panoptic scene graph generation from rich 2d visual scene. *Proceedings of the CVPR*, 2025.
42. Yaoting Wang, Shengqiong Wu, Yuecheng Zhang, Shuicheng Yan, Ziwei Liu, Jiebo Luo, and Hao Fei. Multimodal chain-of-thought reasoning: A comprehensive survey. *arXiv preprint arXiv:2503.12605*, 2025.
43. Hao Fei, Yuan Zhou, Juncheng Li, Xiangtai Li, Qingshan Xu, Bobo Li, Shengqiong Wu, Yaoting Wang, Junbao Zhou, Jiahao Meng, Qingyu Shi, Zhiyuan Zhou, Liangtao Shi, Minghe Gao, Daoan Zhang, Zhiqi Ge, Weiming Wu, Siliang Tang, Kaihang Pan, Yaobo Ye, Haobo Yuan, Tao Zhang, Tianjie Ju, Zixiang Meng, Shilin Xu, Liyu Jia, Wentao Hu, Meng Luo, Jiebo Luo, Tat-Seng Chua, Shuicheng Yan, and Hanwang Zhang. On path to multimodal generalist: General-level and general-bench. In *Proceedings of the ICML*, 2025.
44. Jian Li, Weiheng Lu, Hao Fei, Meng Luo, Ming Dai, Min Xia, Yizhang Jin, Zhenye Gan, Ding Qi, Chaoyou Fu, et al. A survey on benchmarks of multimodal large language models. *arXiv preprint arXiv:2408.08632*, 2024.
45. Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, may 2015. doi:10.1038/nature14539. URL http://dx.doi.org/10.1038/nature14539.
46. 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/.
47. 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.
48. 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.
49. Deli Pei, Huaping Liu, Yulong Liu, and Fuchun Sun. Unsupervised multimodal feature learning for semantic image segmentation. In *The 2013 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–6. IEEE, aug 2013. ISBN 978-1-4673-6129-3. doi:10.1109/IJCNN.2013.6706748. URL http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6706748.
50. Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.
51. 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.),

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 27 June 2025 doi:10.20944/preprints202506.2314.v1

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

52. 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.
53. A. Karpathy and L. Fei-Fei, "Deep visual-semantic alignments for generating image descriptions," *TPAMI*, vol. 39, no. 4, pp. 664–676, 2017.
54. 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.
55. 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.
56. 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.
57. 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.
58. 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.
59. Ling Zhuang, Hao Fei, and Po Hu. Knowledge-enhanced event relation extraction via event ontology prompt. *Inf. Fusion*, 100:101919, 2023.
60. 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.
61. 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.
62. 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.
63. 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.
64. 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.
65. 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.
66. 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.
67. 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.
68. Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*, 2016.
69. 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.
70. D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *ICLR*, 2015.
71. 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.
72. K. Papineni, S. Roukos, T. Ward, and W. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *ACL*, 2002, pp. 311–318.
73. 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.

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 27 June 2025 doi:10.20944/preprints202506.2314.v1

74. Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi:10.18653/v1/N19-1423. URL https://aclanthology.org/N19-1423.
75. Shengqiong Wu, Hao Fei, Leigang Qu, Wei Ji, and Tat-Seng Chua. Next-gpt: Any-to-any multimodal llm. *CoRR*, abs/2309.05519, 2023.
76. 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.
77. 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.
78. Naman Jain, Pranjali Jain, Pratik Kayal, Jayakrishna Sahit, Soham Pachpande, Jayesh Choudhari, et al. Agribot: agriculture-specific question answer system. *IndiaRxiv*, 2019.
79. 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.
80. 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.
81. 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.
82. 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.
83. 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.
84. 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.
85. S. Banerjee and A. Lavie, "METEOR: an automatic metric for MT evaluation with improved correlation with human judgments," in *IEEMMT*, 2005, pp. 65–72.
86. 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.
87. 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.
88. 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.
89. 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.
90. 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.
91. 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.
92. 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.
93. 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.

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 27 June 2025 doi:10.20944/preprints202506.2314.v1

94. 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.
95. 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.
96. 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.
97. P. Anderson, B. Fernando, M. Johnson, and S. Gould, “SPICE: semantic propositional image caption evaluation,” in *ECCV*, 2016, pp. 382–398.
98. 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.
99. 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.
100. 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.
101. 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.
102. Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhutdinov, Richard S Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. *arXiv preprint arXiv:1502.03044*, 2(3):5, 2015.
103. 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.
104. Sanjeev Arora, Yingyu Liang, and Tengyu Ma. A simple but tough-to-beat baseline for sentence embeddings. In *ICLR*, 2017.

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

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.