Emotion-Aware Text Generation for Instructing Audio-Visual Emotion Recognition

Lin Wei[1], Yuanyuan Liu*[1], Kejun Liu[1], Yuxuan Huang[1] and Shaoze Feng[1]
[1]School of Computer Science, China University of Geosciences (Wuhan), Wuhan, China
Linw@cug.edu.cn, liuyy@cug.edu.cn, liukejun@cug.edu.cn, cosinehuang@cug.edu.cn, fengshaoze@cug.edu.cn

***Abstract*—Audio-visual emotion recognition (AVER) is crucial in various AI applications, such as digital humans. However, due to the incomplete and heterogeneity in emotional expression between the audio and visual modalities, challenges remain in achieving effective emotion interaction and fusion. To address these issues, we propose a novel approach, Emotion-Aware Text Generation for Instructing Audio-Visual Emotion Recognition (ETG-AVER). This method generates emotion-Aware text from audio-visual modalities to explicitly capture emotional cues, and utilizes the generated text to guide emotional feature extraction and fusion, thereby improving the performance of AVER. Specifically, ETG-AVER consists of three key components: (1) Emotion-Aware Text Generation module, which first generates Emotion-Aware texts from audio and video modalities, compensating for the limitations of audio-visual modalities in emotional expression; (2) Text-Instruct Multimodal Feature Extraction and Fusion module, which uses the generated Emotion-Aware texts to precisely guide the extraction and fusion of audio-visual features, effectively enhancing cross-modal emotion interaction; (3) Hierarchical Loss Optimization module, which uses hierarchical loss functions to align emotional features across unimodal and multimodal sources, ensuring consistency in both semantic and representational levels, thereby improving cross-modal emotion recognition accuracy. Experimental results demonstrate that ETG-AVER significantly outperforms existing methods in terms of accuracy and fine-grained classification, achieving 60.78% WAR on MAFW and 78.36% WAR on DFEW, validating the effectiveness of our approach in refining emotion information transfer and recognition.**

***Index Terms*—Audio-visual Emotion Recognition, Emotion-Aware Text Generation, Hierarchical Loss**

I. Introduction

Audio-Visual Emotion Recognition (AVER) is an important direction in AI research, with broad applications in fields such as human-computer interaction, digital humans, and intelligent healthcare [1], [2]. Compared to unimodal emotion analysis, AVER simulates human multimodal perception by integrating audio and visual signals, enabling a more comprehensive understanding of human emotions.

Existing audio-visual emotion recognition (AVER) methods primarily employ deep learning techniques to fuse audio and visual features (e.g., Mel-Frequency Cepstral Coefficients(MFCC) [3], spectrograms, facial expressions, and posture) for emotion recognition. To effectively integrate multimodal information, various strategies have been proposed: Ngiam *et al.* [4] introduced Two-Stream Networks, which separately process audio and visual inputs and then merge them via fully connected layers or weighted fusion; Praveen *et al.* [5] leveraged Attention Mechanisms to dynamically allocate weights for capturing important features across modalities; and Hao *et al.* [6] applied Multi-Task Learning to jointly optimize emotion recognition with related tasks for improved model generalization. These methods have significantly enhanced AVER performance, especially in emotion classification tasks, achieving higher recognition rates.

Despite advancements in AVER, two core challenges persist in practical applications. First, the limitations of the audio-visual modalities in emotional expression hinder the transmission of certain emotional cues, making it difficult to capture complete emotional information [7], [8]. Many studies have already demonstrated that, compared to the text modality, the emotional expression in audio-visual modalities is insufficient [9], [10]. Second, the heterogeneity between audio and video modalities complicate their emotion interaction and fusion. Although existing approaches seek to enhance emotion interaction by applying attention mechanisms to weighted fusion of audio-visual features [11], they often fail to adequately capture the emotion-related interaction between the two modalities. Both limitations hinders the overall performance of emotion recognition.

To address the aforementioned challenges, we propose a novel approach, Emotion-Aware Text Generation for Instructing Audio-Visual Emotion Recognition (ETG-AVER). The method generates emotion-aware text from audio-visual modalities, guides precise multimodal feature extraction and fusion using this text, and enhances cross-modal emotion recognition accuracy through hierarchical loss optimization. Specifically, ETG-AVER consists of three core modules: Emotion-Aware Text Generation (ETG), Text-Instruct Multimodal Feature Extraction and Fusion(TMFEF), and Hierarchical Loss Optimization(HLO). To address the limitations of emotional expression in audio-visual modalities, the ETG module creates emotion-focused text from audio-visual data, enhancing emotional richness and improving information transfer. The TMFEF module utilizes the generated text to guide the precise extraction and fusion of audio-visual features, optimizing cross-modal interaction. Lastly, the HLO module refines and aligns emotional features from both unimodal and multimodal sources using hierarchical loss func-

*Corresponding author: Yuanyuan Liu

tions, ensuring effective emotion transfer across modalities and improving recognition accuracy.

The main contributions of our approach are as follows:

- We propose ETG-AVER, a novel approach that generates emotion-aware texts from audio-visual modalities to guide emotion recognition. It alleviates the challenges of incomplete and heterogeneous emotional expression, enhancing fusion effectiveness and significantly improving recognition performance.
- We introduce the ETG module, which generates emotion-aware texts from audio-visual modalities to alleviate the incompleteness in emotional expression. These texts provide explicit emotional cues to instruct subsequent feature extraction and fusion.
- We propose a TMFEF module, where the generated emotion-aware text precisely instructs the extraction and fusion of multimodal features. Through this text-instruction mechanism, we effectively enhance cross-modal emotion interaction, thereby improving overall emotion recognition performance.
- Extensive experiments demonstrate that ETG-AVER outperforms existing methods in emotion recognition accuracy and fine-grained classification, achieving state-of-the-art results on MAFW with a 20.71% relative improvement in WAR over baseline.

II. Related Work

A. Audio-visual Emotion Recognition

Early audio-visual emotion recognition (AVER) methods relied on handcrafted features, such as Mel-frequency cepstral coefficients (MFCC) for audio and the Facial Action Coding System (FACS) for visual modalities. However, these features struggled to capture the complexity of emotional expressions. With the rise of deep learning, CNNs and 3D CNNs have been widely adopted for automatic feature extraction, enhancing recognition accuracy while reducing reliance on manual feature engineering. For instance, Zhang *et al.* [12] employed CNNs and 3D CNNs to generate audio-visual segment features, effectively bridging the emotional gap by integrating multimodal cues within deep models.

To effectively integrate multimodal information, several fusion strategies have been proposed. For instance, Kuhnke *et al.* [13] and Li *et al.* [14] employed Two-Stream Networks, processing audio and visual inputs separately and merging them through fully connected layers or weighted combinations. Ghaleb *et al.* [15] and Praveen *et al.* [5] applied Attention Mechanisms to dynamically assign importance to features across modalities, enhancing the model's focus on key emotional cues. Multi-Task Learning, as demonstrated by Atmaja *et al.* [16] and Hao *et al.* [6], jointly optimizes emotion recognition and related tasks, improving the model's generalization capability.

Despite these advancements, challenges remain in effectively fusing multimodal features and promoting inter-modal collaboration. In recent years, transformer-based models have also been increasingly applied to audio-visual emotion recognition tasks. These models leverage self-attention mechanisms to capture long-range dependencies across modalities, further enhancing the fusion of audio-visual features. Although these methods have significantly improved performance, particularly in emotion classification tasks, challenges such as modality collaboration and cross-modal feature fusion still persist in real-world applications.

B. Text Generation in Multimodal

Early text generation methods primarily relied on manual annotations or rule-based approaches [17]. While these methods were effective, they had significant limitations in terms of flexibility and diversity, leading to text that often fell short in terms of accuracy and breadth, especially in multimodal tasks.

With the rise of deep learning, text generation methods based on pre-trained models (such as BERT [18], GPT-3 [19]) and vision-language models (such as BLIP [20], BLIP2 [21]) have significantly improved the relevance and contextual coherence of the generated text. These models leverage large-scale language and visual data to enhance the quality of the generated content. However, despite these advancements, the quality of generated text still falls short in certain tasks. For instance, in emotion recognition tasks, models like BLIP improve alignment between images and text, but they often fail to capture the nuanced emotions in multimodal content, resulting in text that, while coherent, lacks emotional depth.

While methods such as VPTG [22], which enhances multimodal information fusion and generation in visual dialogue models through visual prompts and knowledge distillation, and VX2TEXT [23], which generates text from video and audio inputs and has shown significant progress in tasks like subtitle generation and question answering, have made strides, challenges remain. In some multimodal tasks, while the generated text becomes more relevant, it still lacks emotional depth and task adaptability, especially when it comes to conveying complex and subtle emotions, and the generated text quality still falls short of fully meeting the demands.

III. Proposed Method

In this section, we present our method, Emotion-Aware Text Generation for Instructing Audio-Visual Emotion Recognition (ETG-AVER). ETG-AVER improves emotion recognition through three core components: (1) the Emotion-Aware Text Generation module (ETG), which addresses the limitations of emotion expression in audio-visual modalities; (2) the Text-Instruct Multimodal Feature Extraction and Fusion module(TMFEF), which enhances modality collaboration; and (3) the Hierarchical Loss Optimization module(HLO), which refines emotion feature transfer across both unimodal and multimodal data, boosting recognition accuracy. The architecture of our approach is illustrated in Figure 1.

A. Emotion-Aware Text Generation

The Emotion-Aware Text Generation module aims to address the incompleteness of emotional expression in audio-

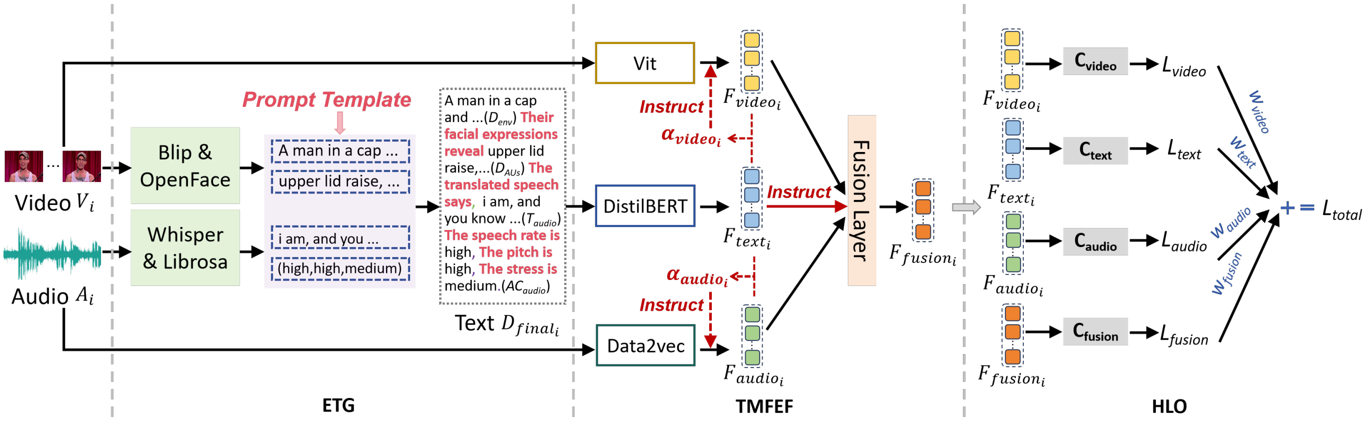


Fig. 1. Framework for Emotion-Aware Text Generation in Instructed Audio-Visual Emotion Recognition. The framework consists of three components: Emotion-Aware Text Generation (ETG), Text-Instruct Multimodal Feature Extraction and Fusion (TMFEF); and Hierarchical Loss Optimization (HLO).

visual modalities. By integrating various audiovisual information, it generates high-quality, emotion-rich textual descriptions, providing deeper semantic information and broader emotional context for emotion analysis, thereby enhancing the model's understanding capabilities.

For the video frame sequence V_i of the i-th sample, we first utilize the BLIP [20] model to generate environment description text D_{env_i}, supplementing the emotional context of the video content. BLIP excels in associating visual information with language, extracting key semantic details. This enables it to provide comprehensive semantic context, overcoming the video modality's limitations in capturing implicit emotional cues, thus enhancing emotion recognition accuracy. Then, to capture more precise and fine-grained emotional expressions, we employ OpenFace [24] to extract facial action units (AUs), which directly reflect an individual's emotional state through distinct facial movements. These AUs are then transformed into structured facial expression descriptions D_{AUs_i} using predefined rules, effectively capturing micro-expressions and their emotional subtleties, thereby providing richer and more nuanced emotional cues for emotion recognition.

For the audio A_i, we apply the Whisper [25] model to transcribe the audio content into linguistic text T_{audio_i}, extracting language-level emotional cues. This transcription enables the capture of semantic information, complementing the emotional expression conveyed through the audio modality. To uncover hidden emotional signals in the audio, we further utilize tools like Librosa [26] to extract acoustic cues AC_{audio_i}, including pitch P_i, rate R_i, and stress S_i:

$$AC_{\text{audio}_i} = \{P_i, R_i, S_i\} \tag{1}$$

These cues reveal the speaker's emotional state and address the limitations of non-verbal emotional expression in the audio modality.

Finally, we use a prompt template T_{template} to combine textual information from both video and audio modalities into a unified description. The template structure is as follows:

> “[D_{env_i}]. Their facial expressions reveal [D_{AUs_i}]. The translated speech says ’[T_{audio_i}]’. The speech rate is [R_i], The pitch is [P_i], and the stress is [S_i].”

This structure merges environment description D_{env_i}, facial expression descriptions D_{AUs_i}, and the linguistic and acoustic elements T_{audio_i} and AC_{audio_i} from the audio into a unified text D_{final_i}:

$$D_{\text{final}_i} = T_{\text{template}}(D_{\text{env}_i}, D_{\text{AUs}_i}, T_{\text{audio}_i}, AC_{\text{audio}_i}) \tag{2}$$

This final description enhances the semantic richness of emotional expression, providing essential support for multimodal emotion recognition tasks.

B. Text-Instruct Multimodal Feature Extraction and Fusion

To enhance cross-modal emotion interaction, we propose the Text-Instruct Multimodal Feature Extraction and Fusion (TMFEF) module. This module utilizes emotion-aware texts to instruct the extraction and fusion of audio-visual features. While audio and video modalities have their strengths, direct concatenation or weighted fusion often fails to fully leverage their complementary nature. The emotion-aware text provides clear emotional cues, enabling the model to more accurately capture and transfer emotion information during the fusion process. By incorporating text instruction, the model effectively coordinates modality relationships, improving emotion transfer precision and recognition accuracy.

We first use DistilBERT [27] to process emotion-aware text D_{final_i}, generating the feature vector $\mathbf{F}_{\text{text}_i}$ that captures emotional cues. These text features not only help the model understand the emotional content but also provide emotional guidance for the audio and video modalities, ensuring that they focus on emotion-relevant information during feature extraction, thereby enhancing emotion recognition.

We use Data2Vec [28] to extract audio features $\mathbf{F}_{\text{audio}_i}$ from the raw audio A_i. To ensure the relevance of the audio features to emotion, emotion-aware text $\mathbf{F}_{\text{text}_i}$ instructs the feature extraction process. Specifically, the emotion-aware text helps the model focus on emotion-relevant components while disregarding irrelevant ones, enhancing the accurate capture and expression of emotional information. We compute the cosine similarity between the text and audio features to quantify the emotional alignment between them:

$$\text{sim}(\mathbf{F}_{\text{text}_i}, \mathbf{F}_{\text{audio}_i}) = \frac{\mathbf{F}_{\text{text}_i} \cdot \mathbf{F}_{\text{audio}_i}}{||\mathbf{F}_{\text{text}_i}||||\mathbf{F}_{\text{audio}_i}||} \tag{3}$$

Based on the similarity score, we compute emotion weights α_{audio_i}, which reflect the contribution of each audio feature to emotion transfer:

$$\alpha_{\text{audio}_i} = \frac{\text{sim}(\mathbf{F}_{\text{text}_i}, \mathbf{F}_{\text{audio}_i})}{\sum_j \text{sim}(\mathbf{F}_{\text{text}_i}, \mathbf{F}_{\text{audio}_j})} \tag{4}$$

These emotion weights α_{audio_i} are used to adjust the importance of the audio features, allowing the model to focus more on emotion-relevant components and ignore irrelevant ones, thereby improving the accuracy of emotion recognition. In this way, the emotion-aware text plays a guiding role in audio feature extraction, helping the model capture and fuse emotional signals more effectively in multimodal emotion recognition.

The video processing follows a similar approach, using Vision Transformer (ViT) [29] for feature extraction to produce the video feature vector $\mathbf{F}_{\text{video}_i}$. Emotion-aware text $\mathbf{F}_{\text{text}_i}$ instructs the extraction of emotion-relevant features. The corresponding emotion weight α_{video_i} is computed and used to adjust the video features.

To capture dynamic emotional interactions across modalities, we utilize the scaled dot-product attention mechanism, where emotion-aware text explicitly instructs the attention process. Unlike traditional methods that rely solely on feature interactions between modalities, our approach leverages emotion-aware text to direct the model's focus to emotion-relevant features, enhancing the interaction and fusion of audio, video, and text features.

For each modality, we apply linear transformations to obtain query, key, and value representations: $\mathbf{Q}_{\text{text}_i} = W_Q\mathbf{F}_{\text{text}_i}$, $\mathbf{K}_{\text{audio}_i} = W_K\mathbf{F}_{\text{audio}_i}$, $\mathbf{V}_{\text{video}_i} = W_V\mathbf{F}_{\text{video}_i}$. Here, emotion-aware text $\mathbf{F}_{\text{text}_i}$ acts as the query, guiding attention between audio and video. Unlike standard attention mechanisms, our approach emphasizes emotion-relevant features in the audio and video through the emotional context provided by the text.

After computing attention scores and normalizing them with softmax, we derive attention weights α_i, which highlight emotional components in the audio and video features. These weights are then used to adjust the video features $\mathbf{V}_{\text{video}_i}$, resulting in the final fused feature vector:

$$\mathbf{F}_{\text{fusion}} = \alpha_i \cdot \mathbf{V}_{\text{video}_i}. \tag{5}$$

This method ensures that emotion-aware text actively guides attention allocation, facilitating effective cross-modal interaction and fusion, and improving emotion recognition performance.

C. Hierarchical Loss Optimization

To enhance the efficiency of emotion feature transfer across unimodal and multimodal data and improve emotion recognition accuracy, we propose a Hierarchical Loss Optimization (HLO) module. This module leverages a multi-level emotion classification strategy, which fine-tunes the transfer of emotional features between modalities, significantly boosting recognition accuracy. By incorporating this approach, we optimize the emotional recognition capabilities of the fused features while strengthening the contribution of each modality, addressing the challenge of insufficient information transfer between fused and unimodal features. This results in a more precise capture and transfer of emotional cues at various levels.

In the HLO module, the fused features and individual modality features are processed separately through distinct classifiers. The fused feature vector $\mathbf{F}_{\text{fusion}}$, obtained via attention-guided fusion, is passed through classifier C_{fusion} for emotion recognition. Simultaneously, the individual modality features $\mathbf{F}_{\text{text}_i}$, $\mathbf{F}_{\text{audio}_i}$, and $\mathbf{F}_{\text{video}_i}$ are input into their respective classifiers C_{text}, C_{audio}, and C_{video} for independent emotion classification tasks.

The outputs of these classifiers, denoted as $\hat{y}_{\text{fusion}}$, $\hat{y}_{\text{text}}$, $\hat{y}_{\text{audio}}$, and $\hat{y}_{\text{video}}$, represent the predicted emotion labels for the fused features and each individual modality, respectively.

This multi-level classification framework allows the model to fully exploit the integrated emotional cues across modalities while also capturing the unique emotional signals within each modality, thereby substantially improving overall emotion recognition performance.

To train the audio-visual emotion recognition (AVER) model effectively, we employ a hierarchical loss function that combines the losses from the fused features and each modality (text, audio, and video). Specifically, we calculate four distinct classification losses corresponding to the fused features and each modality:

$$\mathcal{L}_j = \text{CrossEntropy}(\hat{y}_j, y), \quad j \in \{\text{fusion}, \text{text}, \text{audio}, \text{video}\} \tag{6}$$

The total loss is the weighted sum of these individual losses:

$$\mathcal{L}_{\text{total}} = \sum_j w_j \mathcal{L}_j, \quad j \in \{\text{fusion}, \text{text}, \text{audio}, \text{video}\} \tag{7}$$

Where w_{fusion}, w_{text}, w_{audio}, and w_{video} are the weight coefficients for each loss term, set according to the contribution of each modality to emotion recognition. This hierarchical loss optimization not only enhances the emotion transfer precision of the fused features but also improves the emotion recognition accuracy for audio and video modalities.

IV. Experiments

A. Datasets

We evaluate our method on two widely used multimodal emotion recognition datasets: MAFW [30] and DFEW [31]. The MAFW dataset contains 10,045 video clips with 11 emotion categories. It follows a 5-fold cross-validation setup. The DFEW dataset consists of 16,000 audiovisual clips and poses a 7-class classification task with emotions. Like MAFW, it also follows a 5-fold split for evaluation. Both datasets provide a challenging test for multimodal emotion recognition models.

B. *Experimental setup*

For training, we extract 8 frames per video sequence. The image resolution is set to 224×224, and the batch size is 4. The learning rate is $1e - 5$. We use the AdamW optimizer [32], and the random seed is fixed to 42 for reproducibility.

The experiments are conducted on a single NVIDIA RTX 4090 GPU. We follow the 5-fold cross-validation protocol for both datasets. Evaluation is performed using Unweighted Average Recall (UAR) and Weighted Average Recall (WAR).

C. *Experimental Results and Analysis*

We evaluate the performance of our method on the MAFW and DFEW datasets. To ensure a fair comparison, we follow the experimental setup discussed earlier and report results using five-fold cross-validation. In this section, we present a comparison of our method with several existing state-of-the-art (SOTA) approaches, along with visualizations of the experimental outcomes, including confusion matrices and emotion-aware text generation examples.

1) Performance Comparison: Table I and II present a comparison between our model and other SOTA methods on the MAFW and DFEW datasets.

TABLE I
COMPARISON WITH THE SOTA METHODS FOR AUDIO-VISUAL EMOTION RECOGNITION ON MAFW. THE BEST RESULTS ARE IN BOLD, THE SECOND-BEST RESULTS ARE UNDERLINED.

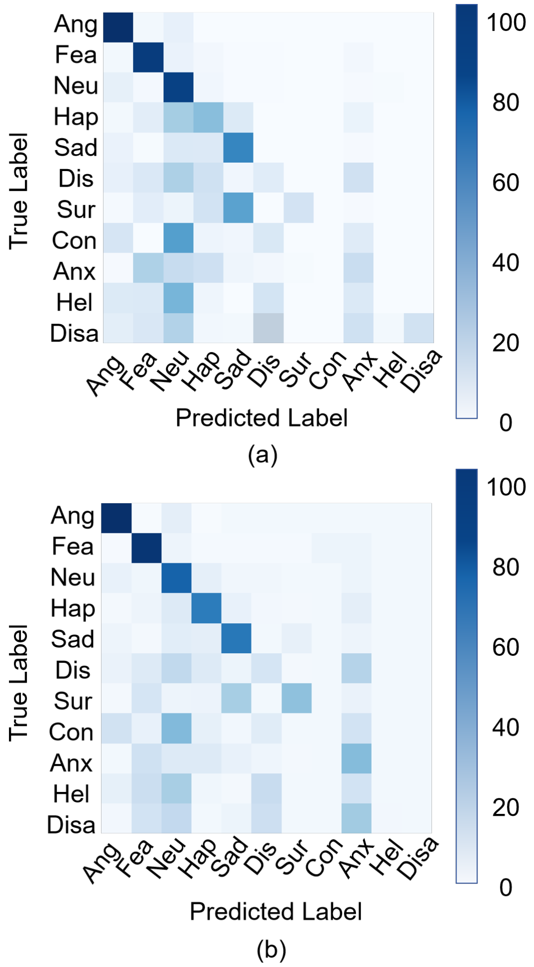
Method	WAR	UAR
C3D+LSTM [30]	44.15	30.47
T-ESFL [30]	48.70	33.35
AMH [33]	48.83	32.98
T-MEP [34]	51.15	37.17
HiCMAE [8]	56.17	42.65
MMA-DFER [35]	58.52	44.25
ETG-AVER (Ours)	**60.78**	**44.27**

TABLE II
COMPARISON WITH THE SOTA METHODS FOR AUDIO-VISUAL EMOTION RECOGNITION ON DFEW. THE BEST RESULTS ARE IN BOLD, THE SECOND-BEST RESULTS ARE UNDERLINED.

Method	WAR	UAR
C3D+LSTM [30]	65.17	53.77
AMH [33]	66.51	54.48
T-MEP [34]	68.85	57.16
HiCMAE [8]	75.01	63.76
UMBEnet [36]	74.83	62.23
MMA-DFER [35]	77.51	**67.01**
ETG-AVER(Ours)	**78.36**	63.83

On the MAFW dataset, our method achieves a WAR of 60.78%, outperforming the second-best approach, MMA-DFER, by 2.26%, with a slightly higher UAR as well. On the DFEW dataset, we further enhance performance, reaching a WAR of 78.36%, surpassing MMA-DFER's 77.51%, further validating the superiority of our approach. These results demonstrate that generating emotion-Aware text from audio-visual modalities to explicitly capture emotional cues, and utilizing the generated text to guide emotion feature extraction and fusion, effectively mitigates the limitations of the audio-visual modalities, significantly improving the model's emotion recognition capabilities. Compared to traditional methods, our ETG-AVER approach not only boosts overall emotion recognition accuracy but also excels in fine-grained emotion classification, particularly on complex datasets such as MAFW.

2) Confusion Matrix: To better illustrate the classification performance of our model, we visualize the confusion matrix on the MAFW dataset, as shown in Figure 2.



(a)

(b)

Fig. 2. Confusion matrices of the Baseline, which only uses a simple concatenation of audio and video features, and a single classifier for classifying the fusion features (a) and our approach (b) on the MAFW dataset. Ang, Fea, Neu, Hap, Sad, Dis, Sur, Con, Anx, Hel, Disa are the abbreviations of the corresponding expression labels. The deeper colors indicate higher accuracy.

Figure 2(a) shows the baseline model results, which use a simple concatenation of audio and video features and a single classifier for fusion feature classification. Figure 2(b) presents the results of our model, which effectively distinguishes between emotion categories with fewer misclassifications, particularly for major emotions. It also performs well in recognizing less-represented categories, such as ”Anxiety,” demonstrating improved accuracy and robustness to data imbalance. However, the model performs worse on “Disgust,” likely due to weak and inconsistent cues across modalities, causing confusion with “Neutral” and “Anxiety.” We view this as a common challenge and plan to address it via contextual modeling and data augmentation.

D. Ablation studies

In this section, we conduct several ablation studies to evaluate the contribution of different components and configurations in our model.

1) Effects of Different Modules: To evaluate the contribution of each key module in our model, we conducted an ablation study by sequentially removing the Emotion-Aware Text Generation(ETG), Text-Guided Multimodal Feature Extraction and Fusion(TMFEF), and Hierarchical Loss Optimization(HLO) modules. We then compared the performance with the complete ETG-AVER model and a baseline model that only uses simple concatenation of audio and video features, with a single classifier for fusion feature classification. The experimental results are summarized in Table III.

TABLE III
THE EFFECT OF DIFFERENT MODULES.

ETG	TMFEF	HLO	WAR	UAR
			50.35	34.24
✓			52.67	35.69
✓	✓		55.45	38.31
✓	✓	✓	**60.78**	**44.27**

As shown in the table III, removing the HLO module causes a performance drop, with WAR decreasing to 55.45% and UAR to 38.31%. Further removing the TMFEF module results in additional declines, with WAR dropping to 52.67% and UAR to 35.69%. Without all modules, the baseline model achieves only 50.35% WAR and 34.24% UAR. These results demonstrate the critical role of each module in boosting emotion recognition performance.

Specifically, the ETG module generates emotion-Aware text to capture emotional cues from audio-visual modalities. The TMFEF module enhances multimodal feature extraction and fusion, while the HLO module optimizes cross-modal emotional feature transfer through hierarchical loss. The synergy of these modules is essential for our model’s superior performance.

2) Effects of Different Fusion Methods in TMFEF: In this section, we aim to explore the impact of different fusion methods on model performance. To do this, we compare our fusion approach, scaled dot-product attention, with alternative methods such as concatenation of feature vectors and additive attention. Table IV presents the results, showing that the scaled dot-product attention method outperforms both concatenation [37] and additive attention [38], achieving the highest WAR of 60.78% and UAR of 44.27%. This demonstrates the effectiveness of using attention mechanisms to better capture the relationships between modalities, leading to improved emotion recognition performance.

TABLE IV
THE EFFECT OF DIFFERENT FUSION METHOD.

Fusion Method	WAR	UAR
Concatenation	56.98	36.64
Additive Attention	57.23	35.98
Scaled Dot-product Attention(Ours)	**60.78**	**44.27**

3) Effects of Different Loss Weight Setting on HLO: To investigate the impact of different weight settings on the classification loss in Equation 7, we conduct an ablation study by adjusting the weights assigned to each classification loss component, including the loss for the fused features and individual modality losses (text, audio, and video). These weight settings are determined based on the contribution of each modality to the overall emotion recognition task. We evaluate three weight configurations as follows:

- **Configuration 1**: Equal weights are assigned to all loss terms (fusion, text, audio, and video). Specifically, each loss is weighted at 1, providing a neutral baseline without emphasizing any particular modality.
- **Configuration 2**: The fusion loss is assigned a higher weight of 2, emphasizing feature fusion over individual modalities, while the losses for text, audio, and video remain at 1 each.
- **Configuration 3 (Ours)**: Weights are adjusted based on modality contribution, with text and fusion assigned higher weights (2 each) due to critical in emotion recognition, while audio and video losses are set to 1.

Table V presents the results of our weight configuration ablation.

TABLE V
THE EFFECT OF DIFFERENT LOSS WEIGHT SETTING.

Loss Weight	WAR	UAR
Configuration 1	53.87	37.24
Configuration 2	57.93	39.66
Configuration 3(ours)	**60.78**	**44.27**

Configuration 1, with equal weights for all modalities, yields lower performance (WAR: 53.87%, UAR: 37.24%), suggesting that a balanced approach is suboptimal for emotion recognition. Configuration 2, which prioritizes feature fusion by increasing the fusion loss weight, shows improved performance (WAR: 57.93%, UAR: 39.66%), emphasizing the value of fusion. However, the best results are achieved in Configuration 3 (our method), where higher weights are assigned to text and fused features, reflecting the text modality’s critical role, leading to significant improvements in WAR (60.78%) and UAR (44.27%). These results confirm the importance of adjusting weights based on each modality’s contribution for optimal performance.

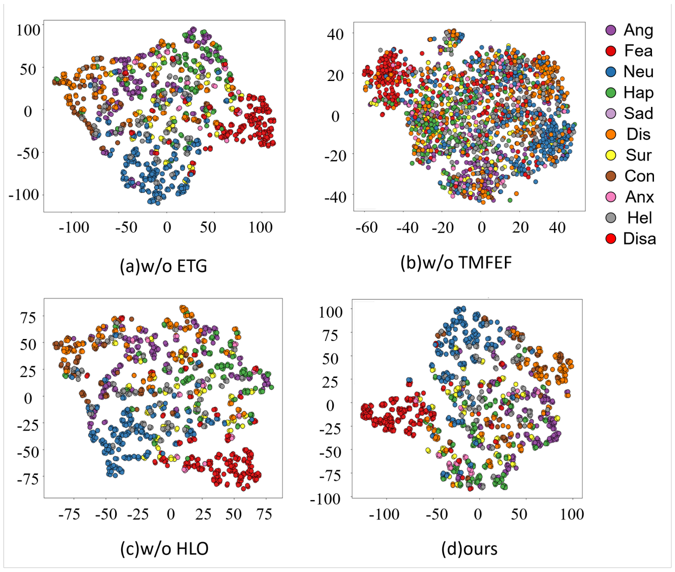
4) Effects of Different Emotion-aware Text Generation Methods: To evaluate the impact of different text content on model performance, we compare three emotion-aware text generation methods: (1) Audio-only: T_{audio} + AC_{audio}, (2) Video-only: D_{env} + D_{AUs}, and (3) Combined Audio-Visual (Ours): D_{env} + D_{AUs} + T_{audio} + AC_{audio}.

TABLE VI
THE EFFECT OF DIFFERENT TEXT CONTENT.

Template	WAR	UAR
Audio-only	52.34	34.18
Video-only	53.67	36.29
Ours	**60.78**	**44.27**

As shown in Table VI, combining text generated from both audio and video modalities significantly improves emotion recognition accuracy. Specifically, using the combined audio-visual text content enables the model to better integrate emotional cues, enhancing both overall recognition accuracy and fine-grained expression detection. This method proves especially effective in complex scenes or emotional expressions, as the fusion of audio and video-generated texts provides richer and more precise emotional context.

5) Visualization of Fusion Feature Distribution Across Ablation Settings: To evaluate the impact of different ablation settings on fusion feature distribution, we perform t-SNE visualization of the model features in a low-dimensional space using the MAFW dataset. The visualizations compare the feature distributions after removing the ETG, TMFEF, and HLO modules with those of the full model.



(a)w/o ETG

(b)w/o TMFEF

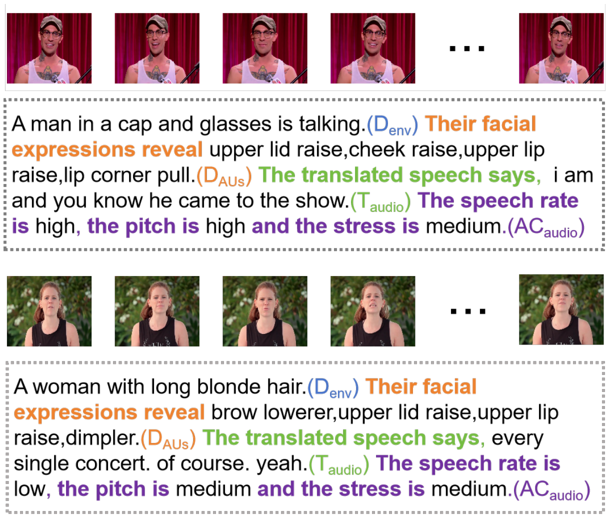
(c)w/o HLO

(d)ours

Fig. 3. t-SNE visualizations of fusion feature distributions for different ablation settings. (a) w/o ETG: Feature distribution after removing the ETG module. (b) w/o TMFEF: Feature distribution after removing the TMFEF module. (c) w/o HLO: Feature distribution after removing the HLO module. (d) Ours: Feature distribution of the complete model with all modules.

The t-SNE results highlight significant variations in feature clustering across ablation settings. Without the ETG module (Figure 3(a)), feature clusters become more dispersed, especially for ”Disgust” and ”Surprise”, demonstrating ETG’s crucial role in enhancing feature cohesion. When the TMFEF module is removed (Figure 3(b)), the feature clusters exhibit greater inter-class overlap and increased intra-class dispersion, indicating diminished feature discriminability and resulting in a noticeable drop in overall performance. This highlights the essential role of TMFEF in facilitating effective multimodal feature fusion. Removing the HLO module (Figure 3(c)) causes concentrated yet scattered clusters, particularly for ”Disgust”, indicating the role of HLO in optimizing feature aggregation. In contrast, the full model (Figure 3(d)) exhibits the clearest clustering and optimal classification performance, confirming the essential contributions of all modules. These findings demonstrate that removing any module disrupts feature clustering and recognition, while the complete model ensures superior performance.

6) Example of Textual Descriptions Generated by ETG: In order to demonstrate the effectiveness of the model in generating text descriptions, we provide examples from the MAFW datasets. These examples showcase the model’s ability to generate text descriptions, as shown in Figure 4.



A man in a cap and glasses is talking.(D$_{\text{env}}$) **Their facial expressions reveal** upper lid raise,cheek raise,upper lip raise,lip corner pull.(D$_{\text{AUs}}$) **The translated speech says,** i am and you know he came to the show.(T$_{\text{audio}}$) **The speech rate is** high, **the pitch is** high **and the stress is** medium.(AC$_{\text{audio}}$)

A woman with long blonde hair.(D$_{\text{env}}$) **Their facial expressions reveal** brow lowerer,upper lid raise,upper lip raise,dimpler.(D$_{\text{AUs}}$) **The translated speech says,** every single concert. of course. yeah.(T$_{\text{audio}}$) **The speech rate is** low, **the pitch is** medium **and the stress is** medium.(AC$_{\text{audio}}$)

Fig. 4. Example predictions of the model generating text descriptions on the MAFW dataset. D_{env} represents the video environment description, D_{AUs} refers to the facial action units descriptive text, T_{audio} represents the transcribed audio content, and AC_{audio} indicates the acoustic cues.

V. CONCLUSION

This paper introduces ETG-AVER, a novel approach for audio-visual emotion recognition that addresses the limitations of emotional expression in audio and visual modalities, as well as their inherent heterogeneity. By generating emotion-Aware text from audio-visual data, capturing emotional cues more effectively, and instructing feature extraction and fusion through the generated text, ETG-AVER enhances audio-visual emotion recognition. The method integrates three key components: Emotion-Aware Text Generation (ETG), Text-Instruct Multimodal Feature Extraction and Fusion (TMFEF), and

Hierarchical Loss Optimization (HLO), which together improve emotion information transfer and recognition accuracy. Extensive experiments on the MAFW and DFEW datasets show significant performance improvements. Future work will focus on integrating additional modalities, refining fusion techniques, and extending ETG-AVER to real-time emotion recognition systems and broader AI applications.

References

[1] R. W. Picard, *Affective computing*. MIT press, 2000.

[2] M. Chen, Y. Zhang, M. Qiu, N. Guizani, and Y. Hao, "Spha: Smart personal health advisor based on deep analytics," *IEEE Communications Magazine*, vol. 56, no. 3, pp. 164–169, 2018.

[3] Z. K. Abdul and A. K. Al-Talabani, "Mel frequency cepstral coefficient and its applications: A review," *IEEE Access*, vol. 10, pp. 122 136–122 158, 2022.

[4] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, A. Y. Ng *et al.*, "Multimodal deep learning." in *ICML*, vol. 11, 2011, pp. 689–696.

[5] R. G. Praveen, E. Granger, and P. Cardinal, "Cross attentional audio-visual fusion for dimensional emotion recognition," in *2021 16th IEEE International Conference on Automatic Face and Gesture Recognition*. IEEE, 2021, pp. 1–8.

[6] M. Hao, W.-H. Cao, Z.-T. Liu, M. Wu, and P. Xiao, "Visual-audio emotion recognition based on multi-task and ensemble learning with multiple features," *Neurocomputing*, vol. 391, pp. 42–51, 2020.

[7] S. Venkatraman, M. Narendra, V. Sharma, S. Malarvannan, A. H. Gandomi *et al.*, "Multimodal emotion recognition using audio-video transformer fusion with cross attention," *arXiv preprint arXiv:2407.18552*, 2024.

[8] L. Sun, Z. Lian, B. Liu, and J. Tao, "Hicmae: Hierarchical contrastive masked autoencoder for self-supervised audio-visual emotion recognition," *Information Fusion*, vol. 108, p. 102382, 2024.

[9] Y. Wang, Y. Shen, Z. Liu, P. P. Liang, A. Zadeh, and L.-P. Morency, "Words can shift: Dynamically adjusting word representations using nonverbal behaviors," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 33, no. 01, 2019, pp. 7216–7223.

[10] M. Chen and X. Li, "Swafn: Sentimental words aware fusion network for multimodal sentiment analysis," in *Proceedings of the 28th international conference on computational linguistics*, 2020, pp. 1067–1077.

[11] R. G. Praveen and J. Alam, "Incongruity-aware cross-modal attention for audio-visual fusion in dimensional emotion recognition," *IEEE Journal of Selected Topics in Signal Processing*, 2024.

[12] S. Zhang, S. Zhang, T. Huang, and W. Gao, "Multimodal deep convolutional neural network for audio-visual emotion recognition," in *Proceedings of the 2016 ACM on International Conference on Multimedia Retrieval*, 2016, pp. 281–284.

[13] F. Kuhnke, L. Rumberg, and J. Ostermann, "Two-stream aural-visual affect analysis in the wild," in *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition*. IEEE, 2020, pp. 600–605.

[14] Y. Li, W. Gan, K. Lu, D. Jiang, and R. Jain, "Aves: An audio-visual emotion stream dataset for temporal emotion detection," *IEEE Transactions on Affective Computing*, pp. 1–14, 2024.

[15] E. Ghaleb, J. Niehues, and S. Asteriadis, "Joint modelling of audio-visual cues using attention mechanisms for emotion recognition," *Multimedia Tools and Applications*, vol. 82, no. 8, pp. 11 239–11 264, 2023.

[16] B. T. Atmaja and M. Akagi, "Multitask learning and multistage fusion for dimensional audiovisual emotion recognition," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 4482–4486.

[17] A. Gatt and E. Krahmer, "Survey of the state of the art in natural language generation: Core tasks, applications and evaluation," *Journal of Artificial Intelligence Research*, vol. 61, pp. 65–170, 2018.

[18] J. D. M.-W. C. Kenton and L. K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of naacL-HLT*, vol. 1, no. 2. Minneapolis, Minnesota, 2019.

[19] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell *et al.*, "Language models are few-shot learners," *Advances in neural information processing systems*, vol. 33, pp. 1877–1901, 2020.

[20] J. Li, D. Li, C. Xiong, and S. Hoi, "Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation," in *International conference on machine learning*. PMLR, 2022, pp. 12 888–12 900.

[21] J. Li, D. Li, S. Savarese, and S. Hoi, "Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models," in *International conference on machine learning*. PMLR, 2023, pp. 19 730–19 742.

[22] M. Zhu, Y. Weng, B. Li, S. He, K. Liu, and J. Zhao, "Knowledge transfer with visual prompt in multi-modal dialogue understanding and generation," in *Proceedings of the First Workshop On Transcript Understanding*, 2022, pp. 8–19.

[23] X. Lin, G. Bertasius, J. Wang, S.-F. Chang, D. Parikh, and L. Torresani, "Vx2text: End-to-end learning of video-based text generation from multimodal inputs," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 7005–7015.

[24] T. Baltrušaitis, M. Mahmoud, and P. Robinson, "Cross-dataset learning and person-specific normalisation for automatic action unit detection," in *2015 11th IEEE international conference and workshops on automatic face and gesture recognition*, vol. 6. IEEE, 2015, pp. 1–6.

[25] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *International conference on machine learning*. PMLR, 2023, pp. 28 492–28 518.

[26] B. McFee, C. Raffel, D. Liang, D. P. Ellis, M. McVicar, E. Battenberg, and O. Nieto, "librosa: Audio and music signal analysis in python." *SciPy*, vol. 2015, pp. 18–24, 2015.

[27] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter," *arXiv preprint arXiv:1910.01108*, 2019.

[28] A. Baevski, W.-N. Hsu, Q. Xu, A. Babu, J. Gu, and M. Auli, "Data2vec: A general framework for self-supervised learning in speech, vision and language," in *International conference on machine learning*. PMLR, 2022, pp. 1298–1312.

[29] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[30] Y. Liu, W. Dai, C. Feng, W. Wang, G. Yin, J. Zeng, and S. Shan, "Mafw: A large-scale, multi-modal, compound affective database for dynamic facial expression recognition in the wild," in *Proceedings of the 30th ACM International Conference on Multimedia*, 2022, pp. 24–32.

[31] X. Jiang, Y. Zong, W. Zheng, C. Tang, W. Xia, C. Lu, and J. Liu, "Dfew: A large-scale database for recognizing dynamic facial expressions in the wild," in *Proceedings of the 28th ACM international conference on multimedia*, 2020, pp. 2881–2889.

[32] I. Loshchilov, "Decoupled weight decay regularization," *arXiv preprint arXiv:1711.05101*, 2017.

[33] S. Yoon, S. Dey, H. Lee, and K. Jung, "Attentive modality hopping mechanism for speech emotion recognition," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 3362–3366.

[34] X. Zhang, M. Li, S. Lin, H. Xu, and G. Xiao, "Transformer-based multimodal emotional perception for dynamic facial expression recognition in the wild," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 34, no. 5, pp. 3192–3203, 2024.

[35] K. Chumachenko, A. Iosifidis, and M. Gabbouj, "Mma-dfer: Multimodal adaptation of unimodal models for dynamic facial expression recognition in-the-wild," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 4673–4682.

[36] X. Mai, J. Lin, H. Wang, Z. Tao, Y. Wang, S. Yan, X. Tong, J. Yu, B. Wang, Z. Zhou *et al.*, "All rivers run into the sea: Unified modality brain-inspired emotional central mechanism," in *Proceedings of the 32nd ACM International Conference on Multimedia*, 2024, pp. 632–641.

[37] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz *et al.*, "Transformers: State-of-the-art natural language processing," in *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, 2020, pp. 38–45.

[38] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *arXiv preprint arXiv:1409.0473*, 2014.