Improving model explainability in dynamic facial expression recognition for hybrid intellectual systems

Victoria Vysotska[1,†], Kirill Smelyakov[2,†], Anastasiya Chupryna[2,†], Anastasiia Kochkina[2,*,†], Ganna Pliekhova[3,†], Anton Naumov[2,†]

[1] *Lviv Polytechnic National University, Stepan Bandera Street, 12, Lviv, 79013, Ukraine*
[2] *National University of RadioElectronics, Nauky Avenue 14, Kharkiv, 61166, Ukraine*
[2] *Kharkiv National Automobile and Highway University, Yaroslava Mudrogo str. 25, Kharkiv, 61002, Ukraine*

Abstract

Face recognition in dynamic real-world conditions presents a challenge due to varying lighting, partial occlusions, and pose variations. While convolutional graph networks (GCNs) and transformer-based architectures achieve high performance in addressing these issues, their black-box nature complicates interpretation. This paper proposes a methodology to enhance the explainability of dynamic face recognition models using graph-based approaches and attention mechanisms for implementing the model to the Hybrid Intellectual System. We introduce visualisation methods for key facial landmarks and frame significance in video sequences through Explainable AI (XAI) techniques, such as Attention Attribution, Feature Ablation, Grad-CAM, and Attention Rollout. Experimental results indicate that the proposed approach can improve model interpretability without compromising accuracy. This research explores a multimodal approach by integrating Llama-based Llasa speech synthesis by combining natural language processing with visual facial expression detection.

Keywords

Dynamic Facial Expression Recognition, Graph Transformers, Explainable AI, Grad-CAM, Hybrid Intellectual Systems, Natural Language Processing, Text to Speech Systems, Conversational Systems

1. Introduction

Facial recognition in real-world dynamic conditions is still one of the most challenging and demanding areas in computer vision. FR has a wide range of practical paths to evolve, from public security technologies to the more detailed analysis of personal emotions and interactions in social robotics. We can see significant progress in recent years; however, the task remains challenging due to the wild dynamic conditions, changing lights, head poses, and the presence of partial occlusions.

In computer vision, in particular, the Facial Expression Recognition (FER) field, Convolutional NNs (CNNs) are used the most frequently for static and dynamic 2D recognition tasks. However, CNN may lose their performance in 3D tasks because of object position changes. Applying Graph Convolutional Networks (GCN) instead for such tasks is reasonable as they effectively capture spatial relationships between facial key points, allowing more precise detection of changes. At the same time, Transformer-based networks [1], initially built for natural language processing tasks, have great potential for temporal sequence processing, capturing dynamic changes with high complexity.

Incorporating NLP-driven models such as Llama-based Llasa into this framework enables a multimodal approach, where facial expressions are not only recognised but also translated into

CLW-2025: Computational Linguistics Workshop at 9th International Conference on Computational Linguistics and Intelligent Systems (CoLInS-2025), May 15–16, 2025, Kharkiv, Ukraine
* Corresponding author.
† These authors contributed equally.
victoria.a.vysotska@lpnu.ua (V. Vysotska); kyrylo.smelyakov@nure.ua (K. Smelyakov); anastasiya.chupryna@nure.ua (A. Chupryna); anastasiia.kochkina@nure.ua (A. Kochkina); plehovaanna11@gmail.com (G. Pliekhova); anton.naumov@nhdsl.com (A. Naumov);
0000-0001-6417-3689 (V. Vysotska); 0000-0001-9938-5489 (K. Smelyakov); 0000-0003-0394-9900 (A. Chupryna); 0009-0000-6679-8746 (A. Kochkina); 0000-0002-6912-6520 (G. Pliekhova); 0009-0007-9100-7978 (A. Naumov)

contextually appropriate speech responses. Llasa integration enables expressive speech synthesis by matching language patterns [27-29] with emotional expressions, ensuring that detected facial expressions can simulate human-like speech. Furthermore, it will enhance affective computing, making human-AI integrations more emotionally aware. Incorporating the scalability of Llama-based architectures, Llasa provides highly accurate speech synthesis that captures not only the lexical content but also the emotional subtext of detected facial expression patterns.

However, networks with a complicated architecture remain "black boxes", which affect more than performance but also end-user trust. This second aspect is crucially vital in specific domains such as medicine, security, and personal data processing.

The issue of interpretability has gained increasing importance due to the growing interest in transparency and explainability in AI-driven decision-making. Users and developers must understand which features and frames are decisive in a model's predictions. This is why integrating explainable AI (XAI) approaches into face recognition models is a key direction for enhancing their practical value.

Human conversation is naturally multimodal: meaning is encoded not only in text but also in facial expression, vocal prosody and gesture. These paralinguistic features colour literal words, attitude and emotional tone that are essential for smooth flow and mutual understanding. Dialogue systems and social agents that rely on text alone, therefore, miss a substantial portion of the communicative channel, leading to responses that can feel tone-deaf or robotic. Multimodal affective language understanding seeks to fuse verbal and non-verbal cues so that artificial agents can interpret and respond with emotionally congruent behaviour, bringing machine interaction closer to human conversational norms.

This study aims to develop and integrate specialised XAI methods into graph-based and Transformer architectures to improve recognition accuracy and significantly enhance the interpretability of decision-making processes. Specifically, the focus is on analysing and visualising the most relevant facial landmarks and key video frames that substantially impact the final classification outcome. In addition, this research explores the integration of SpatioTemporal Graph Transformer (STGT) with the Llasa text-to-speech synthesis model [4] to create a hybrid multimodal system, ensuring an emotionally adaptive AI system [13].

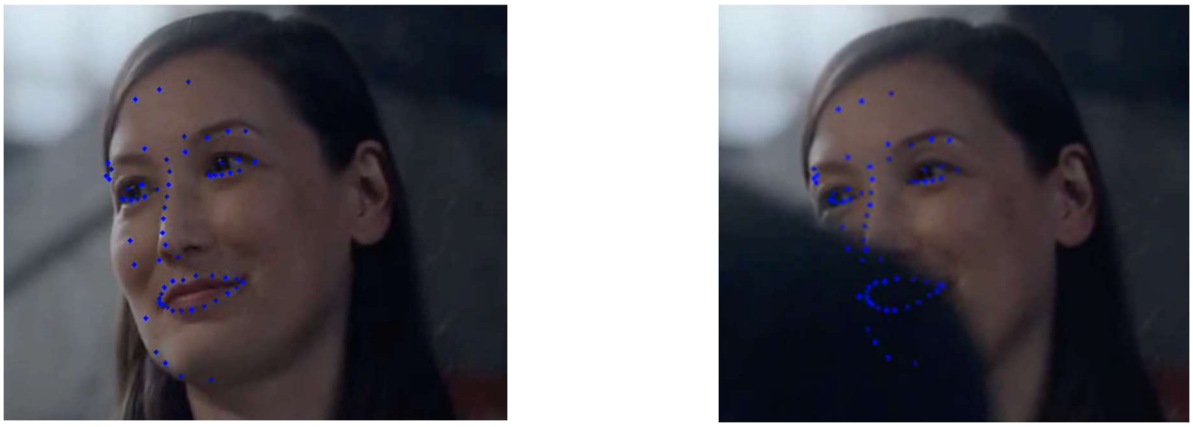
2. Related works

2.1. Graph Methods for Dynamic Facial Expression Recognition

Graph-based approaches leverage facial landmark points to construct a structured representation of the face, where each node corresponds to a specific coordinate in space. This method enables efficient modelling of spatial and temporal relationships between facial regions, which is crucial for dynamic facial expression recognition (DFER) [15]. Unlike traditional 2D representations, increasing the dimensionality from 2D to 3D enhances the ability to capture subtle depth variations in facial expressions, leading to more accurate and robust emotion classification. One of the most effective frameworks for real-time 3D facial landmark extraction is MediaPipe [5], developed by Google. It defines a 3D face model using 468 key points, which are detected and tracked across video sequences. The MediaPipe Face Landmark Model normalises the X and Y coordinates within a range of 0 to 1. At the same time, the Z-coordinate is estimated relative to the X-axis using a perspective projection camera model, with values ranging between -1 and 1. This approach ensures a consistent and stable representation of depth, which is critical for analysing microexpressions and subtle facial movements in dynamic sequences. MediaPipe shows a strong ability to detect landmarks on frames with interferences, as shown in Figure 1.

Given its high computational efficiency and real-time processing capabilities even on mobile devices, MediaPipe serves as a strong foundation for integrating graph-based methods into Graph Neural Networks (GNNs) [14] and Transformer models for DFER [15]. The framework's ability to accurately capture 3D facial dynamics in real-time makes it highly suitable for graph-based facial

representation [25], enabling more expressive feature extraction while maintaining a balance between performance and accuracy. Additionally, the integration of 3D landmarks with graph-based models allows for improved spatial reasoning, helping the model better understand facial deformations over time.



(A) **(B)**

Figure 1: MediaPipe Face Mesh [5] detected landmarks on MAFW [3] sample (A) landmarks detected without interferences; (B) landmarks detected with interferences

2.2. Implementing Transformer-based networks for DFER

Transformer-based networks have recently gained attention for Dynamic Facial Expression Recognition (DFER) due to their effectiveness in modelling long-range dependencies in sequential data. Unlike traditional Convolutional neural networks (CNNs), which primarily focus on spatial feature extraction, or Long Short-Term Memory networks (LSTMs) and Recurrent neural networks (RNNs), which capture temporal patterns but often struggle with long-term dependencies, Transformers use self-attention mechanisms to process facial expressions across multiple frames holistically.

DFER is increasingly proficient at analysing spatiotemporal features [2] from video sequences. The ViViT [23] model improves on traditional CNN-RNN hybrid approaches by employing factorised attention mechanisms to directly extract both spatial and temporal representations, thereby eliminating the need for recurrent layers. These abilities allow us to pick up subtle patterns of changes in facial expression over time.

ViViT works by dividing video frames into patches, which are embedded in a high-dimensional feature space. These patch embeddings then pass through spatiotemporal self-attention layers, allowing the model to effectively learn spatial dependencies alongside temporal variations. It enables accurate detection of subtle facial movements (patterns), including micro-expressions.

Additionally, ViViT's scalability to more extended video sequences makes it well-suited for real-world applications where emotions change dynamically. By integrating pose-invariant and occlusion-aware learning strategies it ensures robust performance under varying conditions. Recent studies [6] show that ViViT outperforms traditional CNN-LSTM models and other Transformer models when dealing with pose variations, lighting changes, and expression intensity changes, highlighting its promise in the field of deep learning-based affective computing.

In DFER, each video sequence can be represented as a temporal-spatial graph, with facial landmarks serving as nodes and their relationships evolving. A Spatial Transformer within the network learns to focus on critical facial regions through landmark-level attention. At the same time,

a Temporal Transformer captures the sequential dependencies between frames, allowing the system to detect subtle transitions in expressions.

This architecture is particularly beneficial for addressing challenges such as occlusions, pose variations, and fine-grained emotional cues, ultimately enhancing recognition accuracy.

2.3. Explainable AI (XAI) for DFER

Explainable AI (XAI) in dynamic facial expression recognition (DFER) enhances model transparency and fosters trustworthiness by tackling challenges such as temporal dependencies, pose variations, and micro-expressions. Traditional facial expression recognition models, such as RNNs and CNNs, often perform as black-box classifiers, making their predictions difficult to interpret. To mitigate this issue, techniques like LIME and SHAP [24] are employed to identify influential facial features, while Grad-CAM [7] visualises the key facial regions that contribute to the classifications.

DFER requires models capable of processing sequences over time. Transformers, which utilise self-attention mechanisms [2], allow models to focus on critical frames, thereby improving interpretability. Graph-based approaches represent facial landmarks as nodes in a graph neural network (GNN) and employ attention mechanisms to dynamically highlight important regions. In addition, adaptations of Grad-CAM for temporal-spatial graphs can generate heatmaps over time, enhancing the explainability of the model. Prototype-based methods, such as ProtoPNet, provide case-based reasoning, and Contrastive Explanation Methods (CEMs) help distinguish subtle differences in expressions.

Evaluating XAI in DFER involves human-in-the-loop studies, faithfulness tests, and ensuring alignment with psychological theories of emotion perception. However, challenges remain in interpreting temporal features, mitigating bias, and ensuring real-time explainability for applications in healthcare and surveillance. Future research should aim to develop scalable, real-time XAI techniques to further enhance the transparency and effectiveness of DFER models.

3. Methods and Materials

3.1. Dataset and Preprocessing

The MAFW dataset served as the foundation for data preprocessing. Initially, video files were processed using MediaPipe Face Mesh, which facilitated the extraction of a comprehensive set of 468 key points representing facial landmarks. To optimise the performance of the Transformer model, a deliberate reduction of these key points was implemented, narrowing them down to a more manageable 68.

These selected key points were systematically organised into distinct datasets according to the number of frames present in each video segment. As a result, three separate datasets were meticulously curated for videos containing up to 50, 100, and 150 frames, which collectively amounted to a substantial total of 8,120 video entries. However, it is worth noting that not every video was paired with corresponding class labels, prompting an adjustment of the final count to 7,706 videos.

The data structure prepared for model training is provided in **Figure 2.**

3.2. Model Architecture

The proposed SpatioTemporal Graph Transformer (STGT) is designed for dynamic facial recognition by capturing both spatial and temporal dependencies in facial landmark sequences.

- The Landmark Embedding Layer converts raw 3D landmark coordinates into a higher-dimensional representation. It employs a linear transformation to map the input format of (B, T, N, 3) into an embedding space of (B, T, N, embed_dim).

- Spatial Transformer Block applies multi-head self-attention to model the spatial dependencies across facial landmarks within a single frame. It processes the data in the format (B, N, T, embed_dim) to allow landmarks (nodes) to attend to one another. Additionally, it uses Layer Normalization and Feedforward Networks to enhance the representations.
- Temporal Transformer Block captures sequential dependencies between frames using stacked Transformer Encoder layers. The input format (B, T, embed_dim) is passed through these layers to model long-range dependencies. It also incorporates gradient tracking hooks to facilitate Grad-CAM-based interpretability.
- Classification Head representations are aggregated using Global Average Pooling over time. The pooled features are then passed through a fully connected layer for final classification.

	clip_name	T_frame	label	x0	y0	z0	x1	\
0	00019.json	0	5	0.398745	0.203495	-0.041064	0.425867	
1	00019.json	1	5	0.398172	0.199947	-0.037238	0.425291	
2	00019.json	2	5	0.397061	0.199594	-0.037346	0.424351	
3	00019.json	3	5	0.396368	0.201792	-0.037634	0.423466	
4	00019.json	4	5	0.394634	0.200031	-0.037633	0.421918	

	y1	z1	x2	...	z64	x65	y65	z65	\
0	0.199946	-0.039751	0.450780	...	0.000267	0.421870	0.523082	0.001159	
1	0.195821	-0.035886	0.450128	...	-0.001719	0.419704	0.522611	-0.000866	
2	0.195394	-0.035746	0.449256	...	-0.002179	0.417977	0.522486	-0.001349	
3	0.197278	-0.036459	0.448318	...	-0.001542	0.419543	0.521060	-0.000600	
4	0.195525	-0.036299	0.446926	...	-0.001748	0.417088	0.522978	-0.000863	

	x66	y66	z66	x67	y67	z67
0	0.414129	0.514805	0.003303	0.408782	0.507868	0.006442
1	0.412001	0.514819	0.001314	0.406835	0.508386	0.004602
2	0.410188	0.514938	0.000820	0.404951	0.508738	0.004112
3	0.411840	0.513410	0.001594	0.406664	0.507003	0.004830
4	0.409384	0.515294	0.001322	0.404251	0.508980	0.004598

[5 rows x 207 columns]

Figure 2: Data formatted and refactored into solid dataset

3.3. XAI implementation

In dynamic facial expression recognition (DFER) using Graph Transformer models, explainability is essential for understanding how the model interprets spatial-temporal facial landmarks to infer emotions. Several techniques can enhance the interpretability and trustworthiness of the model, including Attention Attribution, Feature Ablation, Grad-CAM, and Attention Rollout.

Attention Attribution [19] is the method which helps identify which facial landmarks contribute the most to the model's decision by analysing the self-attention weights in the Transformer layers. Since the Graph Transformer processes 68 facial landmarks as nodes, Attention Attribution can highlight which regions of the face (e.g., eyebrows, mouth, eyes) are most relevant to specific expressions over time.

Feature Ablation systematically removes or masks specific input features (facial landmarks) to assess their impact on model predictions. This technique can be applied to subsets of facial nodes or temporal frames to determine whether certain facial regions or time steps are crucial for emotion classification. For instance, it can help verify whether jaw movements or subtle eye changes are more significant for detecting emotions like anger or sadness.

Grad-CAM is widely used in CNN-based vision [6] models. Still, it can also be adapted to Transformer-based architectures by visualising the importance of different regions in an input sequence. In a Graph Transformer model for DFER, Grad-CAM can generate heatmaps over

the facial graph nodes, indicating which parts of the face influence the model's classification at different time steps.

Attention Rollout aggregates attention maps across multiple layers of the Transformer, providing insights into how low-level local attention in early layers evolves into global contextual attention in deeper layers. This technique is instrumental in hybrid models combining local graph-based learning and global Transformer attention, helping to analyse whether the model progressively captures short-term facial micro-expressions before forming high-level temporal patterns.

3.4. Metrics

The evaluation of the SpatioTemporal Graph Transformer for dynamic facial expression recognition is carried out using multiple performance metrics that access both classification accuracy and model interpretability. Integrated metrics ensure that the system effectively recognises emotions and provides insights into which spatial (facial landmarks) and temporal (frame-based changes) patterns contribute the most to predictions.

For training stability, we will use Cross-Entropy Loss with class weights, as the dataset may contain an imbalanced distribution of emotions. Class weights are computed dynamically to ensure balanced learning.

The weighted cross-entropy loss used to handle class imbalance was calculated using the formula:

$$L_{CE} = -\sum_{i=1}^{C} w_i y_i log_{\hat{y}_i} \tag{1}$$

where

- C is the number of emotion classes,
- w_i is the weight assigned to class i,
- y_i is the ground truth label (one-hot encoded),
- $\hat{y}_i$ is the predicted probability for class i.

The loss function is designed to penalise incorrect predictions based on class imbalance, ensuring that smaller-class emotional patterns are learned effectively.

For model performance evaluation, we use training and validation loss as baseline standards. The training process logs epoch-wise loss values to monitor the model's convergence. A learning rate scheduler adjusts the learning rate dynamically when validation loss stagnates, which prevents overfitting. We use the confusion matrix to achieve a detailed breakdown of model predictions versus actual labels. It highlights misclassification patterns, helping refine the feature extraction process. The matric is plotted using seaborn heatmaps for visual analysis of prediction distributions.

For classification robustness, we use Grad-CAM [7] for Spatial and Temporal attention analysis. Grad-CAM is used to visualise which facial landmarks are most influential in classification. We will create two types of heatmaps for spatial Grad-CAM (which highlights key facial features that contribute to predictions) and temporal Grad-CAM (which identifies critical time frames where emotion transitions occur). These visualisations provide explainability, making the model's decisions more interpretable.

Grad-CAM generates attention heatmaps by computing the gradient of the largest class score with respect to the feature maps. The Grad-CAM activation for a given location (x, y) is:

$$L_{Grad-CAM}^{(c)}(x, y) = ReLU\left(\sum_{k} \alpha_k^c A_k(x, y)\right) \tag{2}$$

where

- $A_k(x, y)$ is the activation map of the k-th convolutional feature map,
- α_k^c is the importance weight computed as:

$$\alpha_k^c = \frac{1}{Z} \sum_{x,y} \frac{\partial S^c}{\partial A_k(x, y)}, \tag{3}$$

- $S^c$is the predicted score for class c,
- Z is the total number of spatial locations.

For temporal Grad-CAM, this process is extended over time by computing gradients across sequential frames.

For model selection and optimisation, we provide checkpoints with the best validation loss. The system automatically saves the best model based on validation loss. A model is only saved if it improves validation loss and maintains a reasonable training loss (>0.4) to prevent premature convergence.

By integrating classification metrics, loss analysis, confusion matrix visualisation and interpretability techniques, we ensure a comprehensive evaluation of the hybrid SpatioTemporal Graph Transformer and Llasa speech synthesis system. The combination listed above ensures not only accuracy but the trustworthiness and interpretability of the system, making it suitable for real-world affective computing applications.

3.5. Integrating Model to the Hybrid Intellectual System

To build a Hybrid Intellectual System, we will integrate the SpatioTemporal Graph Transformer (STGT) at the perception front end alongside a speech-to-text transformer that supplies live transcripts. The resulting affect stream and transcripts are fed to the dialogue-reasoning layer, which drives the Llama-based Llasa speech synthesis model [4], which relies on language processing [9] to ensure natural and contextually appropriate emotional outcomes. Using self-attention mechanisms, the STGT assigns importance to key patterns, identifying expressions (anger, happiness, sadness, surprise, fear, etc.) and pushes small messages of emotion label and confidence to the dialogue manager that updates 20 times per second. The manager stores the latest label in its state slot for user emotion and conduct, prompting the underlying language model to produce an empathic response. For instance, if the user's emotion is Sad, it asks the language model for a gentle, caring response. For Angry, it changes to calm and solution-focused wording.

The chatbot will be fine-tuned on videos containing real-world dialogue samples between 2 people, where it will learn some behaviour patterns and then refine them with human feedback to provide addressing – not mirroring response style.

The finished sentence, along with matching emotional labels, is mapped to linguistic prosody [22] and translated into linguistic attributes such as pitch variation, speech rate and pause insertions.

Llasa incorporates text transformers (Llama-based architecture) to interpret the semantic meaning behind the generated or predefined speech content [21]. The speech tokeniser ensures that the emotion detected in the dialogue manager message aligns with the intonation and rhythm of the synthesised speech.

Each emotion-labelled sentence undergoes a transformation to match the phonetic and prosodic attributes of expressive speech. For example, for angry expressions, the speech sounds sharp and has increased volume.

Llasa employs vector quantisation (VQ) codecs to convert text-based tokens into expressive speech waveforms. Using Process Reward Models (PRMs), the system interactively refines speech synthesis by adjusting articulation, tone and rhythm to align with both visual emotional cues and linguistic context [10]. The final speech is temporally synchronised with an emotion label

corresponding to the user's facial expressions, ensuring that spoken words and tone appear together naturally.

4. Experiments and results

4.1. Model definition and training

Original data was merged into a tabular format, describing each frame for video. An embedding layer transformed these landmarks into higher-dimensional vectors. Facial landmarks were treated as nodes in a graph for each frame, utilising multi-head attention to model spatial dependencies. Temporal block processed frame-level embeddings, capturing temporal relationships across frames. The classifier head aggregated the features to predict expression categories from 11 classes. Initially, the model was trained on a dataset of 50 frames to establish baseline accuracy. Then, the dataset was expanded to include sequences of 100 frames to improve the model's capability to handle extensive temporal contexts. The last dataset containing the most extended sequences of 150 frames was used for training to optimise the model's capability to handle extensive temporal contexts. The model's output metrics are provided in Table 1. Memory-efficient training methods and computational optimisations were applied to ensure efficiency and enhance performance.

Table 1
GCN + Transformer Model Trained

classes	Precision	Recall	F1-Score	Support
1	0.78	0.79	0.78	543
2	0.90	0.93	0.91	566
3	0.92	0.61	0.76	545
4	0.88	0.85	0.86	561
5	0.99	0.97	0.98	572
6	0.92	0.89	0.90	564
7	0.83	0.96	0.89	566
8	0.90	0.80	0.85	558
9	0.86	0.92	0.88	562
10	0.94	0.99	0.96	565
11	0.98	1.00	0.99	563
Accuracy			0.90	6165
Macro avg	0.89	0.89	0.89	6165
Weighted avg	0.91	0.90	0.91	6165

The accuracy of the trained model is shown in **Figure 3**.

4.2. XAI methods implementation

To highlight the aspects that the model uses in its decision-making process, we implemented XAI methods for a more precise analysis. Attention attribution was implemented by extracting attention weights directly from the model's Multi-Head Attention layers. Within the Spatial Transformer Block, we obtained attention weights for each frame, revealing how much importance the model assigned to each facial landmark during decision-making. Similarly, in the Temporal Transformer Block, attention weights were extracted across frames to identify key temporal segments influencing the classification outcomes. The attention weights are shown as heatmaps of facial landmarks, clearly indicating which points are most important for the accurate classification of facial expressions. It provides a clearer understanding of the decision-making process in the model.

A systematic feature ablation strategy was applied to test the significance of features. Specific subsets (subgroups) of facial landmarks were gradually removed, and the resulting change in

classification accuracy was observed. This study allowed us to determine which facial features contributed the most to the model's decision-making process. By quantifying these effects, we gain confidence in the interpretability and robustness of the model.

Grad-CAM was integrated to further enhance interpretability. Grad-CAM enabled visualisation of the gradients flowing into the last convolutional layer (adapted to our transformer-based approach), highlighting specific facial landmarks and regions significantly influencing classification outcomes, as shown in **Figure 4**. It allowed us to visually verify the robustness of the model's attention mechanisms and to confirm the correct identification of critical facial areas linked to specific emotional expressions.

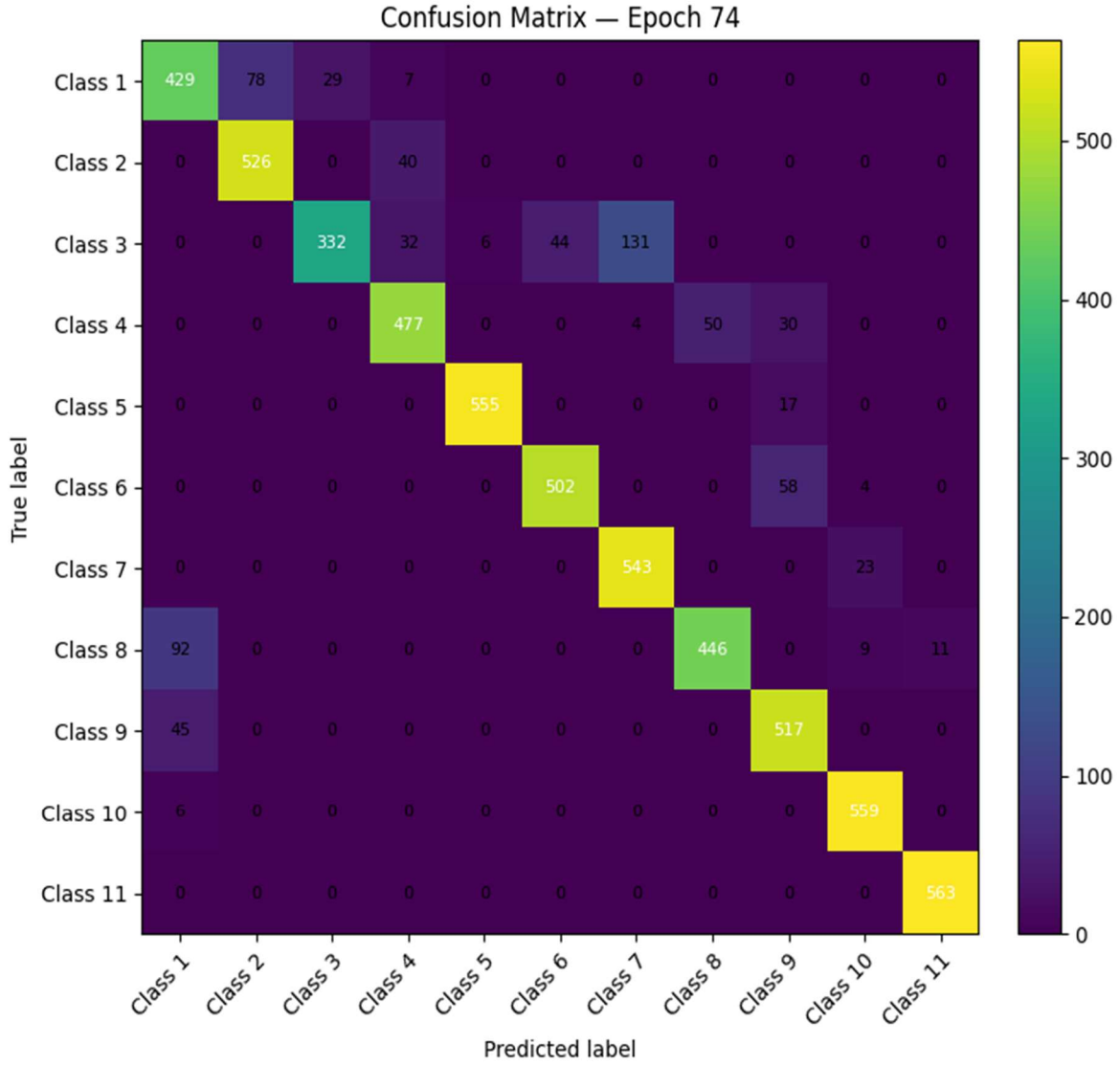
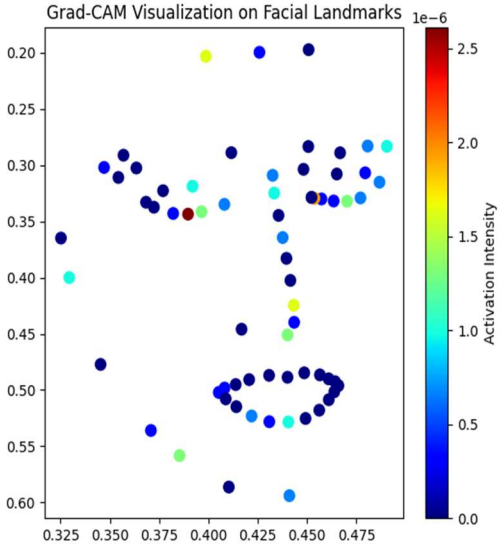


Figure 3: Confusion matrix for STGT model after 74 training epochs

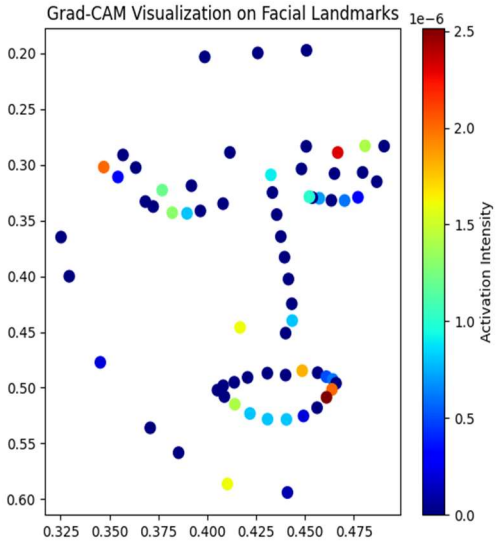
Temporal Grad-CAM implemented in the project provides visualisation of changes in the model's attention for different frames in sequences. Figure 5 visualises the epoch 10 model, which shows unrelated diffuse attention, but the 60th epoch model gathers more extended periods where attention is high.

Layer-wise Relevance Propagation [18] was used extensively to quantify and visualise the contributions of individual landmarks and specific frames toward classification decisions. By propagating relevance scores from the output back to the input features, LRP [17] enabled a detailed analysis of each landmark's influence across time, facilitating the understanding of how temporal dynamics affect the model's predictions.

Taken together, these complementary XAI techniques offer a triangulated view of the model's decision-making process. Insights gleaned from the heatmaps and relevance scores feed directly into an iterative training loop, guiding hyper-parameter tuning and data-augmentation choices that further sharpen both accuracy and transparency.

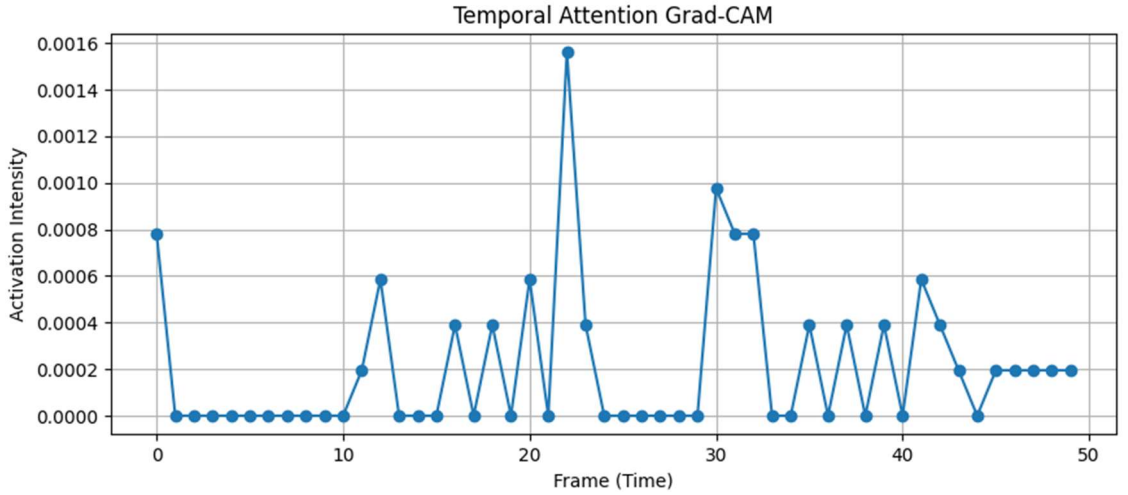


(A)

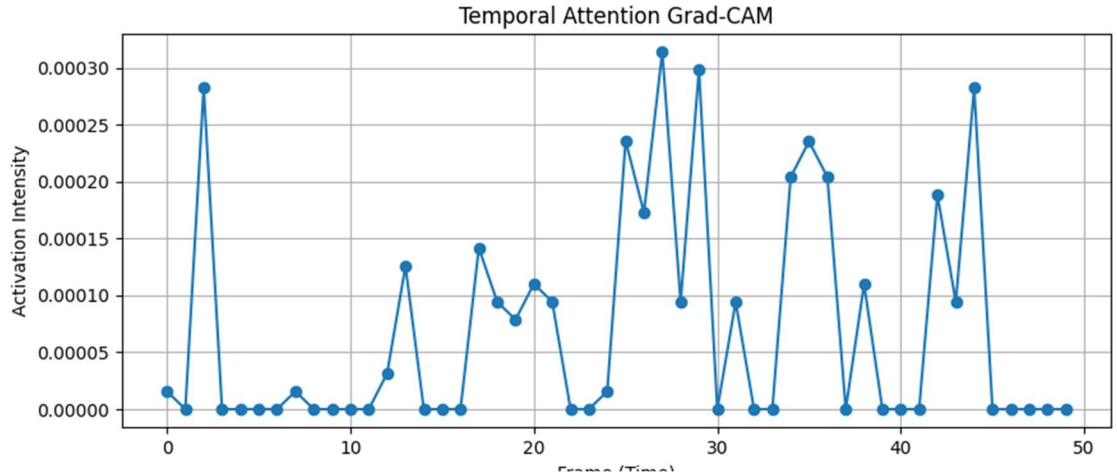


(B)

Figure 4: (A) Grad-CAM shows important nodes on 10th epoch; (B) Grad-CAM shows important nodes on 60th epoch – targeted class – 2 - Disgust.



(A)



(B)

Figure 5: Temporal Grad-CAM on (A) 10th epoch; (B) 60th epoch for targeted class 2 - Disgust

5. Discussion and future research

The results of this study demonstrate the effectiveness of the SpatioTemporal Graph Transformer (STGT) in improving the explainability of Dynamic Facial Expression Recognition (DFER) while maintaining high classification accuracy. By combining graph-based facial landmark processing with transducer-based temporal modelling, the system effectively captures both spatial dependencies and temporal changes in facial expressions. In addition, the integration of methods for increasing explainability, such as Grad-CAM, attention attribution, and feature ablation, provides essential insights into the model's decision-making process, increasing interpretability and confidence in AI-based face recognition systems.

In plans is experimenting with YOLO[11] for multiple face recognition and comparing its capability with MediaPipe in dynamic facial expression recognition. MediaPipe offers lightweight landmark detection in real-time, and YOLO[12] object detection may provide higher precision in challenging conditions. This comparison will help determine what method works best and if integrating both would be beneficial.

Recent advancements in transformer-based architectures, such as ViViT (Video Vision Transformer), have shown that self-attention mechanisms can significantly improve performance by modelling long-range dependencies more effectively than CNN-RNN hybrids. ViViT segments video

frames into patch embeddings and processes both spatial and temporal information cohesively, removing the necessity for recurrent structures. However, ViViT does not clearly encode the structural relationships between facial landmarks, which limits its interpretability in deep fake emotion recognition (DFER) tasks.

A significant point of the proposed method is its ability to highlight facial landmarks that make the most impact and keyframes in video sequences, making sure that classification is not purely "black-box" on output. Increased interpretability is extremely important in healthcare, security, and human-computer interaction applications, where understanding how and why a model makes a specific prediction is as important as its accuracy. Moreover, using attention-based explanations allows the model to be adjusted based on real-world changes such as lighting conditions such as lightning conditions, occlusions and head position.

Despite these advancements, several challenges remain. We will move from independent modules to a unified STGT – ASR perception stack feeding a dialogue manager that, in turn, drives Llasa TTS. The incorporation of Llasa would enable multimodal emotional AI in the conversational system, where detected facial expressions dynamically affect emotionally expressive speech output. Future work will be focused on implementing real-time synchronisation between facial expression recognition and speech generation [26], ensuring that spoken emotions match detected facial cues.

Additionally, current Grad-CAM-based explainability techniques primarily focus on spatial attention rather than temporal dependencies in facial expressions. Future improvements should explore temporal Grad-CAM visualisations to better understand how the model tracks emotion transitions over time, while language-side XAI – via token-level attention rollouts will expose why the bot chooses an appropriate expression in response.

Another perspective way for work includes increasing the diversity of data used for training. To advance beyond clip-level emotion tagging, future work should train the pipeline on long-form dialogue video datasets that align face video, raw speech and full transcripts across multi-sentence turns. Resources such as MELD, IEMOCAP, CMU-MOSEI/MOSI and SEWA supply precisely this alignment, allowing STGT to model avoiding facial affect, ASR to capture prosodic cues, and the dialogue manager to learn how emotions ebb and flow throughout an extended exchange. Leveraging such material will equip the agent to maintain affective context over multiple turns and to generate responses that are not merely reactive but emotionally coherent within the broader conversation.

Facial expressions and their dynamic patterns can differ across age groups, ethnicity and social context. The model should be trained more in a broader range of samples to ensure its robustness and clearer classification ability. For real-world applications, it is mandatory to ensure that potential biases are decreased to a minimum before deploying. Future enhancement of the model could explore lightweight transformer architectures or efficient attention mechanisms such as sparse attention or low-rank adaptations to reduce processing overhead while maintaining accuracy.

6. Conclusions

This study introduces a comprehensive solution for Dynamic Facial Expression Recognition (DFER) by combining SpatioTemporal Graph Transformer (STGT) methods with Explainable AI (XAI) techniques and Llama-based Llasa speech synthesis. The framework captures spatial relationships among facial landmarks and the temporal dynamics of facial movements, achieving accurate and interpretable emotion classification.

To enhance model interpretability, the research utilises advanced XAI methods like Grad-CAM, Attention Attribution, and Feature Ablation. These techniques allow for clear visualisation of the facial features and temporal segments that influence the model's decisions, addressing the challenge of interpretability in AI-driven facial recognition systems and providing valuable insights for users and developers.

Moreover, the integration of NLP-driven speech synthesis via Llasa TTS substantially enriches the system's multimodal interaction capabilities. This enhancement facilitates the synchronised expression of recognised emotions is converted into matching prosody, so the agent voices its

response that fits the user's affect through natural and expressive speech outputs. Such multimodal integration is particularly impactful for advancing applications within human-computer interaction, affective computing, accessibility tools, and other interactive AI-driven environments.

The study demonstrates the integration of an event-driven pipeline – STGT for facial expression affect, a transformer ASR for live transcripts, a dialogue manager that fuses the two, and Llasa TTS for prosody-controlled speech that robust visual emotion recognition with expressive speech synthesis. It also identifies specific areas requiring further enhancement. Particularly, achieving real-time synchronisation between user expressional input and corresponding voice-moduled output remains challenging due to computational constraints and latency issues.

Research shows that transformer-based architectures surpass traditional CNNs and LSTMs in DFER tasks. However, challenges remain with dataset biases, occlusion handling, and head pose variations. Future research should focus on optimising real-time model performance, refining XAI methodologies for better temporal explainability, and diversifying training datasets to ensure fair performance across demographic groups.

In conclusion, this study contributes to emotionally responding to AI by enhancing both model accuracy and interpretability. The proposed hybrid system establishes a robust and adaptable foundation for future exploration and developments in multimodal emotion recognition systems, intelligent interactive agents, and a range of practical affective computing applications.

7. Declaration on Generative AI

The authors have not employed any Generative AI tools.

8. References

[1] Vaswani, A., et al.: Attention is All you Need. In: Advances in Neural Information Processing Systems 30. (2017). https://doi.org/10.48550/arXiv.1706.03762

[2] Wang, Zerui and Yan Liu. "STAA: Spatio-Temporal Attention Attribution for Real-Time Interpreting Transformer-based Video Models." (2024) URL: https://www.semanticscholar.org/reader/6bfa663955410c4f59d5b9bbdd29f8c36670c463

[3] Yuanyuan Liu, Wei Dai, Chuanxu Feng, Wenbin Wang, Guanghao Yin, Jiabei Zeng, and Shiguang Shan. 2022. MAFW: A Large-scale, Multimodal, Compound Affective Database for Dynamic Facial Expression Recognition in the Wild. In Proceedings of the 30th ACM International Conference on Multimedia (MM' 22), October 10–14, 2022, Lisboa, Portugal. ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3503161.3548190 - https://mafw-database.github.io/MAFW/

[4] Ye, Zhen, Xinfa Zhu, Chi-min Chan, Xinsheng Wang, Xu Tan, Jiahe Lei, Yi Peng, Haohe Liu, Yizhu Jin, Zheqi Dai, Hongzhan Lin, Jianyi Chen, Xingjian Du, Liumeng Xue, Yunlin Chen, Zhifei Li, Lei Xie, Qiuqiang Kong, Yi-Ting Guo and Wei Xue. "Llasa: Scaling Train-Time and Inference-Time Compute for Llama-based Speech Synthesis." (2025). Doi: https://doi.org/10.48550/arXiv.2502.04128

[5] Sánchez-Brizuela, G., et al.: Lightweight real-time hand segmentation leveraging MediaPipe landmark detection. Virtual Reality. (2023). https://doi.org/10.1007/s10055-023-00858-0.

[6] Byzkrovnyi, O., Savulioniene, L., Smelyakov, K., Sakalys, P., Chupryna, A. Comparison of Potential Road Accident Detection Algorithms for Modern Machine Vision System, Vide. Tehnologija. Resursi - Environment, Technology, Resources, 2023, 3, pp. 50–55. doi: https://doi.org/10.17770/etr2023vol3.7299

[7] Wang, S., Zhang, Y.: Grad-CAM: Understanding AI Models. Computers, Materials & Continua. 76(2), 1321–1324 (2023). Doi: https://doi.org/10.32604/cmc.2023.041419.

[8] Chen, J., et al.: Attention-Based Multimodal Multi-View Fusion Approach for Driver Facial Expression Recognition. IEEE Access. 1 (2024). Doi: https://doi.org/10.1109/access.2024.3462352

[9] K. Smelyakov, D. Karachevtsev, D. Kulemza, Y. Samoilenko, O. Patlan and A. Chupryna, "Effectiveness of Preprocessing Algorithms for Natural Language Processing Applications," 2020 IEEE International Conference on Problems of Infocommunications. Science and Technology (PIC S&T), Kharkiv, Ukraine, 2020, pp. 187-191, doi: 10.1109/PICST51311.2020.9467919.

[10] Smelyakov, K., Chupryna, A., Darahan, D., Midina, S. Effectiveness of modern text recognition solutions and tools for common data sources, CEUR Workshop Proceedings, 2021, 2870, pp. 154–165, https://ceur-ws.org/Vol-2870/

[11] R. P. Narwaria, A. Ahirwar, A. K. Prajapati, A. Kumar and A. K. Tiwari, "Smart Object Detection Using ESP32-CAM Based on YOLO Algorithm," *2024 Second International Conference on Intelligent Cyber Physical Systems and Internet of Things (ICoICI)*, Coimbatore, India, 2024, pp. 817-820, doi: 10.1109/ICoICI62503.2024.10696374.

[12] Velychko, D., et al.: Image Preprocessing and YOLO Architectures for Enhanced Small and Slow-Moving Object Detection. In: 2024 IEEE Western New York Image and Signal Processing Workshop (WNYISPW), Rochester, NY, USA, 8 Nov 2024, pp. 1–4. IEEE (2024). https://doi.org/10.1109/wnyispw63690.2024.10786503.

[13] Song, T., et al.: MPED: A Multimodal Physiological Emotion Database for Discrete Emotion Recognition. IEEE Access. 7, 12177–12191 (2019). https://doi.org/10.1109/access.2019.2891579.

[14] Zhao, Q., et al.: Density Division Face Clustering Based on Graph Convolutional Networks. In: 2022 26th International Conference on Pattern Recognition (ICPR), Montreal, QC, Canada, 21–25 Aug 2022. IEEE (2022). https://doi.org/10.1109/icpr56361.2022.9956670.

[15] Tian, Y., Li, M., Wang, D.: DFER-Net: Recognising Facial Expression In The Wild. In: 2021 IEEE International Conference on Image Processing (ICIP), Anchorage, AK, USA, 19–22 Sep 2021. IEEE (2021). https://doi.org/10.1109/icip42928.2021.9506770.

[16] Chumachenko, K., Iosifidis, A., Gabbouj, M.: MMA-DFER: MultiModal Adaptation of unimodal models for Dynamic Facial Expression Recognition in-the-wild. In: 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Seattle, WA, USA, 17–18 June 2024, pp. 4673–4682. IEEE (2024). https://doi.org/10.1109/cvprw63382.2024.00470.

[17] Bhati, D., et al.: Neural Network Interpretability with Layer-Wise Relevance Propagation: Novel Techniques for Neuron Selection and Visualization. In: 2025 IEEE 15th Annual Computing and Communication Workshop and Conference (CCWC), Las Vegas, NV, USA, 6–8 Jan 2025, pp. 00441–00447. IEEE (2025). https://doi.org/10.1109/ccwc62904.2025.10903721.

[18] Jung, Y.-J., Han, S.-H., Choi, H.-J.: Explaining CNN and RNN Using Selective Layer-Wise Relevance Propagation. IEEE Access. 9, 18670–18681 (2021). https://doi.org/10.1109/access.2021.3051171.

[19] Lei, S., et al.: Watch the Speakers: A Hybrid Continuous Attribution Network for Emotion Recognition in Conversation With Emotion Disentanglement. In: 2023 IEEE 35th International Conference on Tools with Artificial Intelligence (ICTAI), Atlanta, GA, USA, 6–8 Nov 2023. IEEE (2023). https://doi.org/10.1109/ictai59109.2023.00133.

[20] Li, H., Miao, S., Feng, R.: DG-FPN: Learning Dynamic Feature Fusion Based on Graph Convolution Network For Object Detection. In: 2020 IEEE International Conference on Multimedia and Expo (ICME), London, United Kingdom, 6–10 July 2020. IEEE (2020). https://doi.org/10.1109/icme46284.2020.9102838.

[21] Ahn, Y., Chae, J., Shin, J.W.: Text-to-Speech With Lip Synchronization Based on Speech-Assisted Text-to-Video Alignment and Masked Unit Prediction. IEEE Signal Processing Letters. 1–5 (2025). https://doi.org/10.1109/lsp.2025.3537949.

[22] Fu, Z., et al.: Emotion recognition based on multimodal physiological signals and transfer learning. Frontiers in Neuroscience. 16 (2022). https://doi.org/10.3389/fnins.2022.1000716.

[23] Sun, G., Lian, Z.: Deepfake Video Detection Based on the Decomposition of Spatial-Temporal Attention Mechanism in ViViT. In: 2024 IEEE International Symposium on Parallel and Distributed Processing with Applications (ISPA), Kaifeng, China, 30 Oct–2 Nov 2024, pp. 1629–1634. IEEE (2024). https://doi.org/10.1109/ispa63168.2024.00221.

[24] Oveis, A.H., et al.: Explainability In Hyperspectral Image Classification: A Study of Xai Through the Shap Algorithm. In: 2023 13th Workshop on Hyperspectral Imaging and Signal Processing: Evolution in Remote Sensing (WHISPERS), Athens, Greece, 31 Oct–2 Nov 2023. IEEE (2023). https://doi.org/10.1109/whispers61460.2023.10430776.

[25] Zheng, H., et al.: A separable spatial-temporal graph learning approach for skeleton-based action recognition. IEEE Sensors Letters. 1–4 (2024). https://doi.org/10.1109/lsens.2024.3475515.

[26] S. Das and D. Das, "Natural Language Processing (NLP) Techniques: Usability in Human-Computer Interactions," *2024 6th International Conference on Natural Language Processing (ICNLP)*, Xi'an, China, 2024, pp. 783-787, doi: 10.1109/ICNLP60986.2024.10692776.

[27] J. Francis and M. Subha, "An Overview of Natural Language Processing (NLP) in Healthcare: Implications for English Language Teaching," *2024 8th International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC)*, Kirtipur, Nepal, 2024, pp. 824-827, doi: 10.1109/I-SMAC61858.2024.10714890.

[28] V. Vysotska, N. Sharonova, A. Chupryna, M. Shirokopetleva, O. Dolhanenko, S. Smelyakov, Research of methods for image sharpness evaluation in photos of people, CEUR Workshop Proceedings, Vol-3664, 2024, pp. 255–272.

[29] Q. Zhang, Z. Wang, D. Zhang, W. Niu, S. Caldwell, T. Gedeon, Y. Liu, Z. Qin, "Visual Prompting in LLMs for Enhancing Emotion Recognition," Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 4484-4499, doi: 10.18653/v1/2024.emnlp-main.257